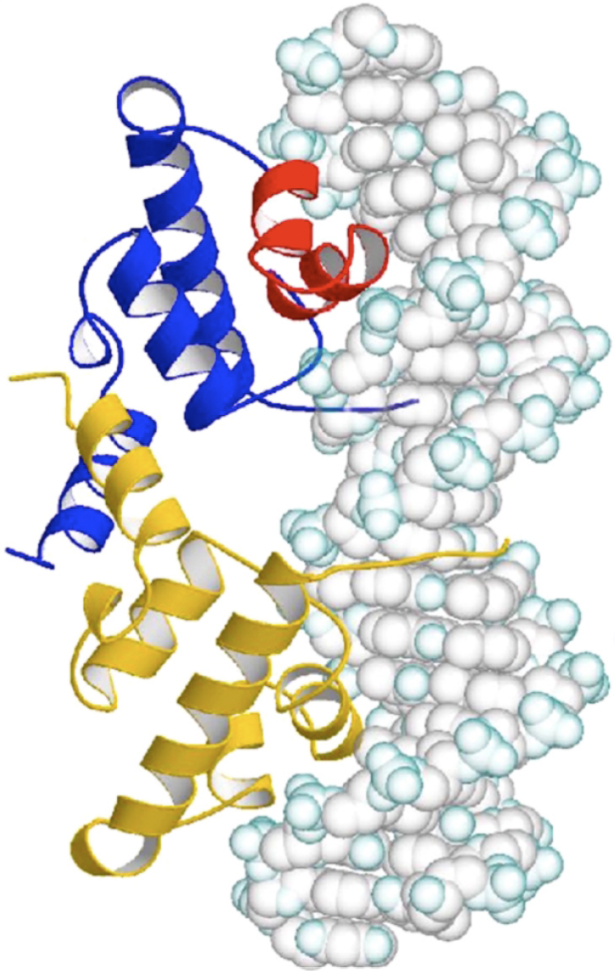
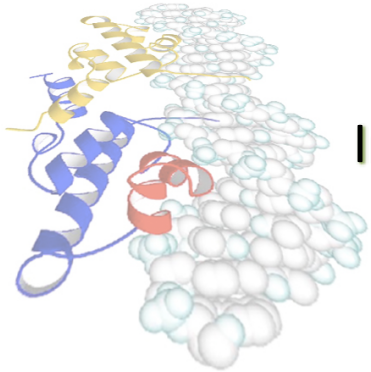




Systems Biology of Transcriptional Regulation

Varodom Charoensawan

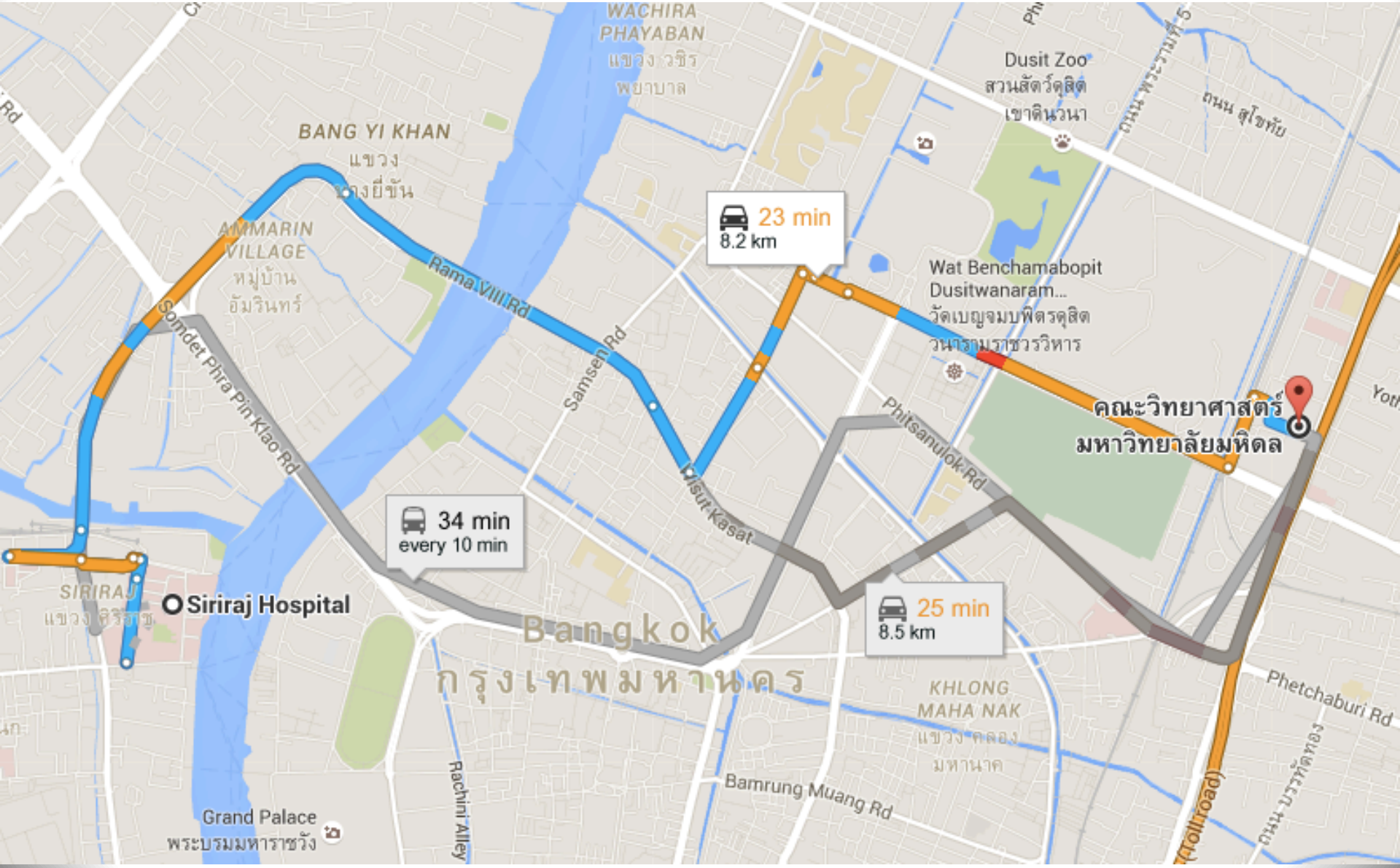
Department of Biochemistry, Faculty of Science

Integrative Computational Bioscience (ICBS) Center

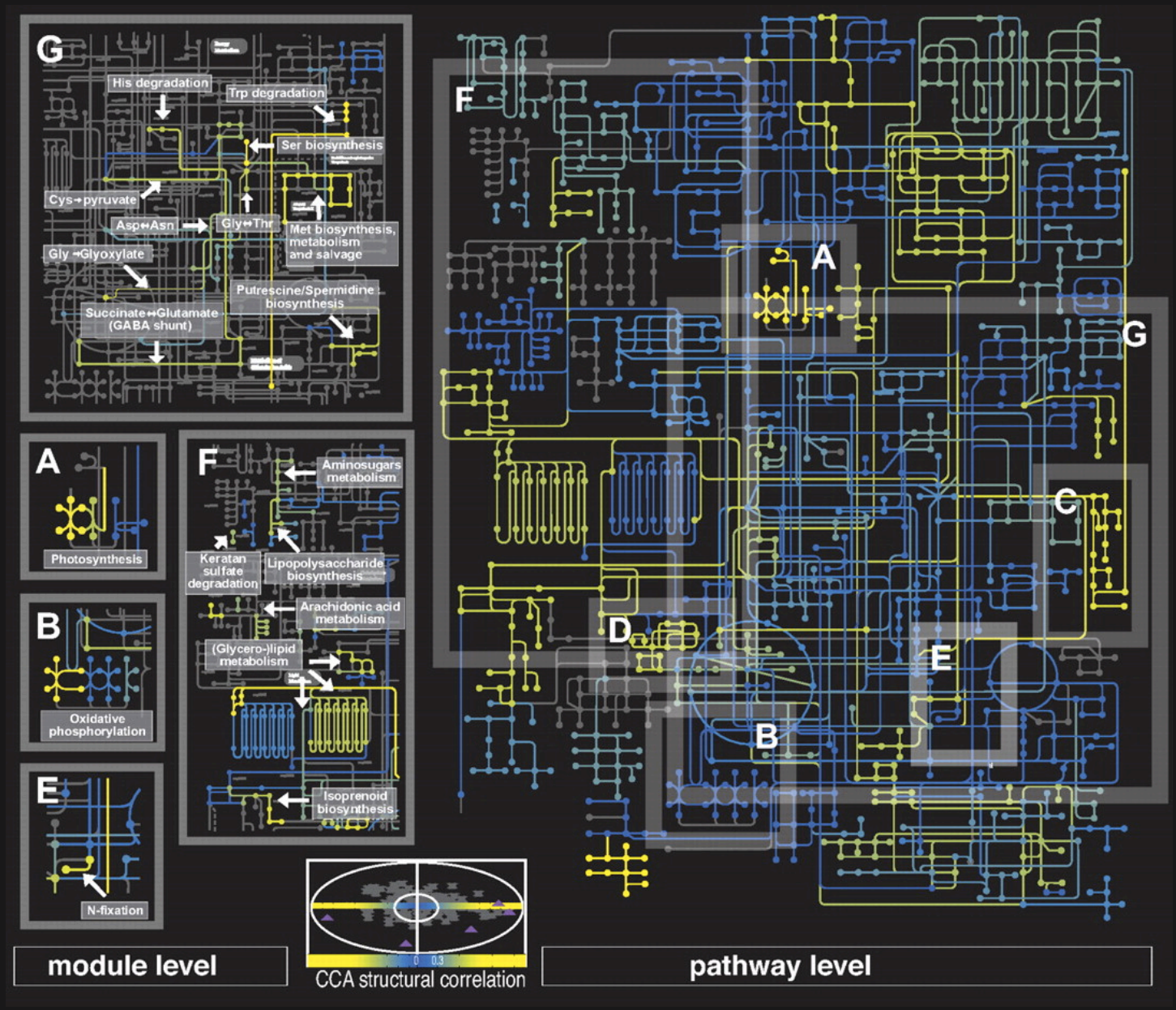
MAHIDOL UNIVERSITY

Wisdom of the Land

How to get from A to B?



A complete map of an organism?



Gianoulis *et al. PNAS* (2008) 106(5): 1374–1379,

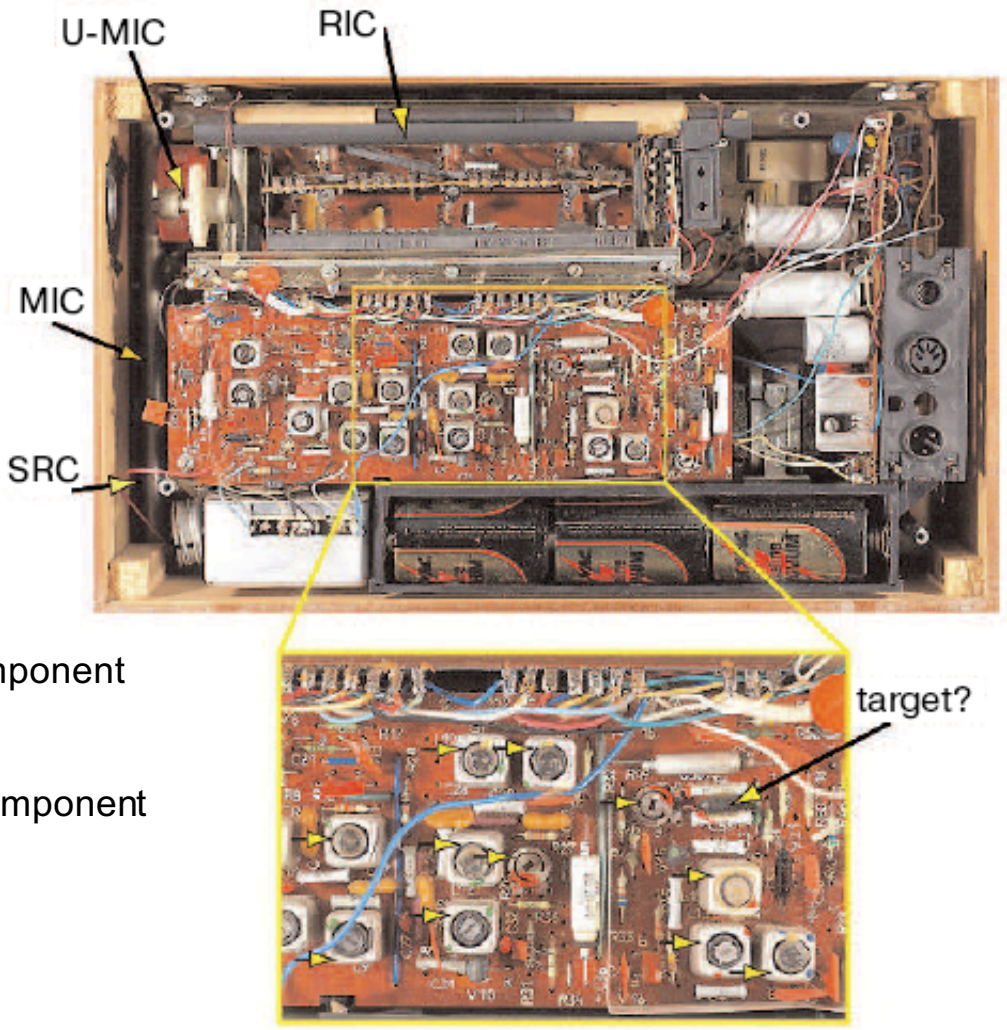
"Systems Biology"

- Investigates links between multiple genes, proteins at one time, rather than one individual gene, protein (genome-wide, large-scale and –omics data)

- Usually involved combination of different techniques, tools: wet-lab experiment + computational analyses

Creating a biological "map" of an organism

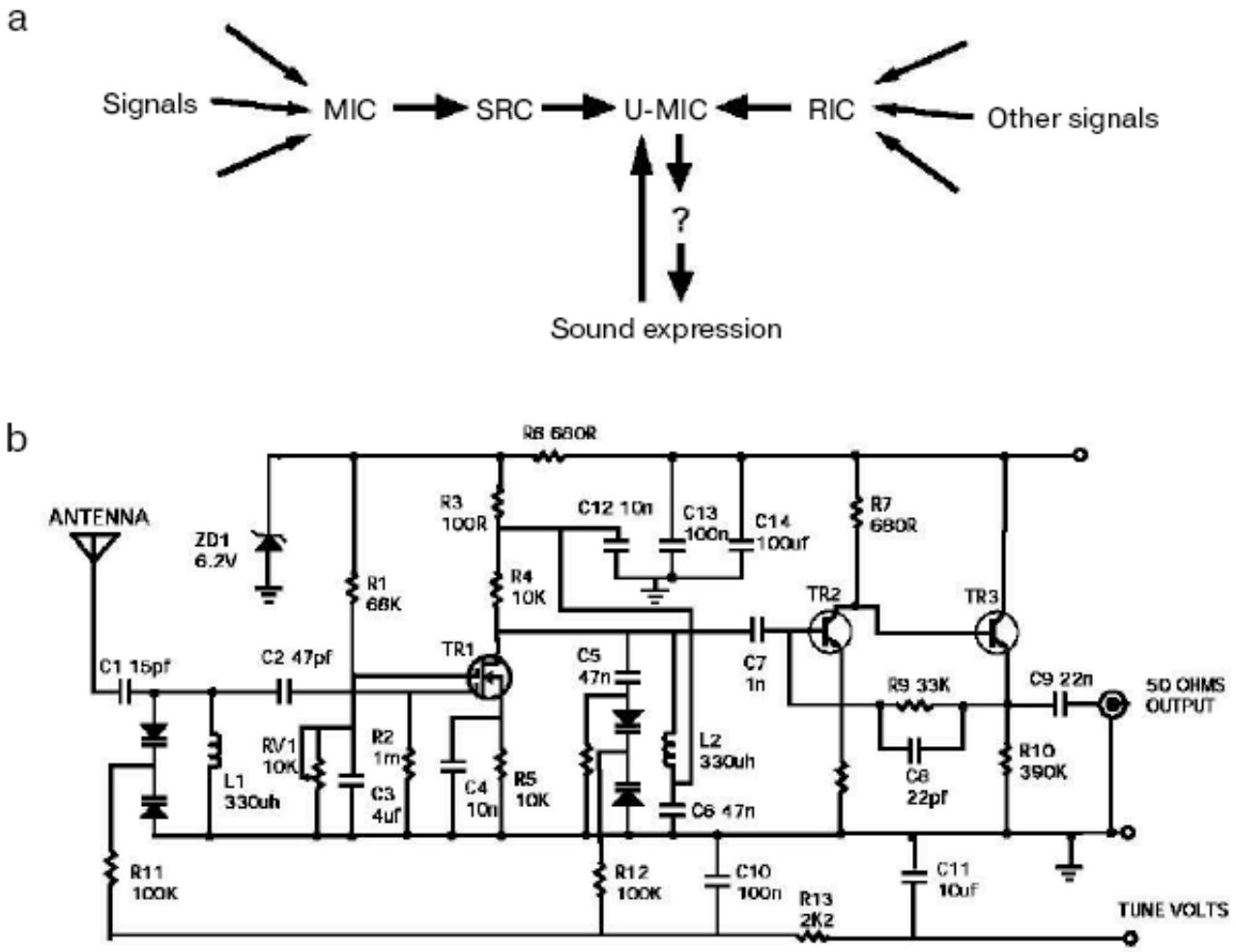
Can a biologist fix a radio? (1)



- Serendipitously Recovered Component
- Really Important Component
- Most Important Component
- Undoubtedly Most Important Component

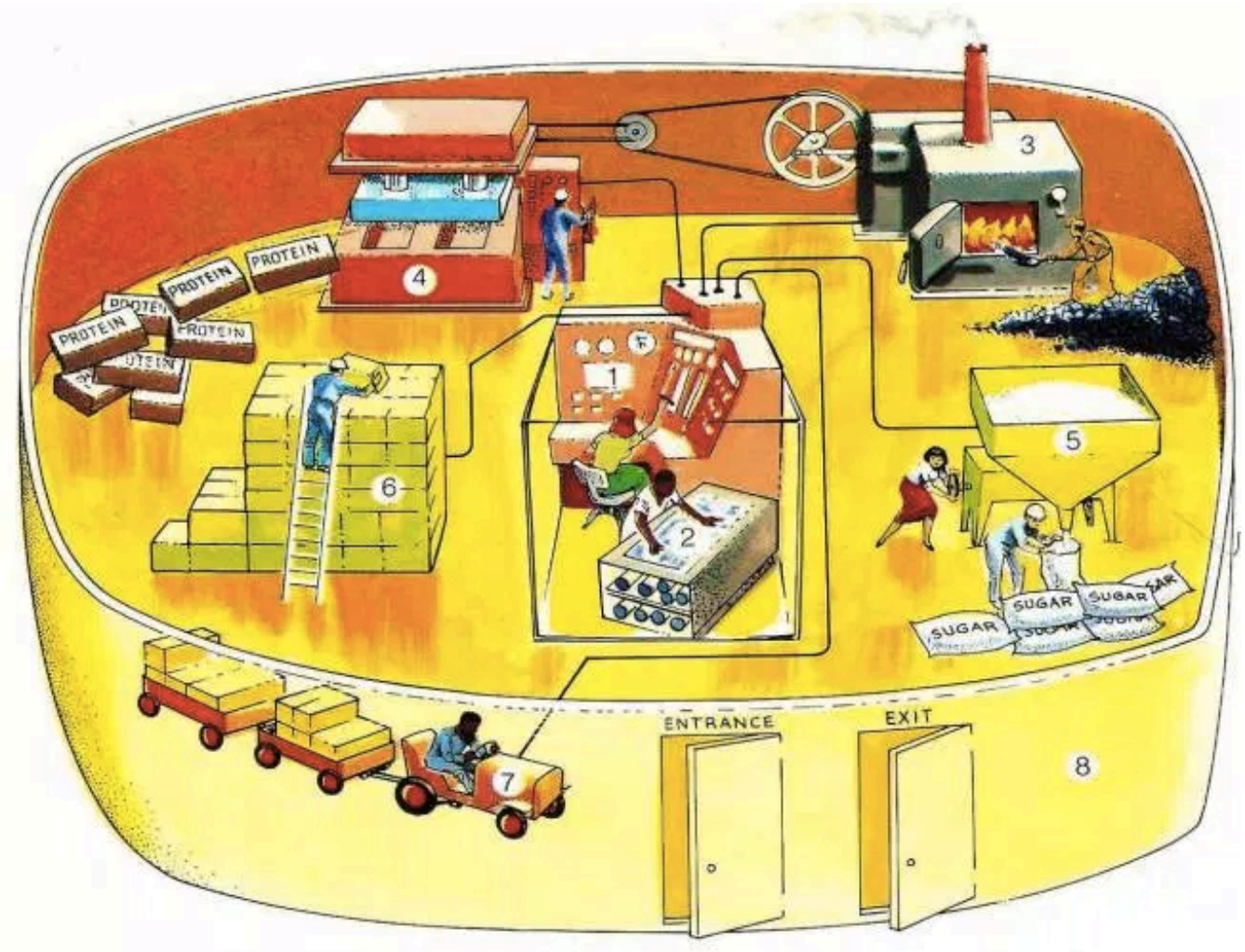
Lazebnik Y., Biochemistry (Moscow) 2004

Can a biologist fix a radio? (2)



Lazebnik Y., Biochemistry (Moscow) 2004

Cell Factory



1. Nucleus (control center)

2. Chromosomes (file cabinet, contains all the information)

3.Mitochondria (power house)

4. Ribosomes and RER (assembly line, work benches)

5. Chloroplast (energy source)

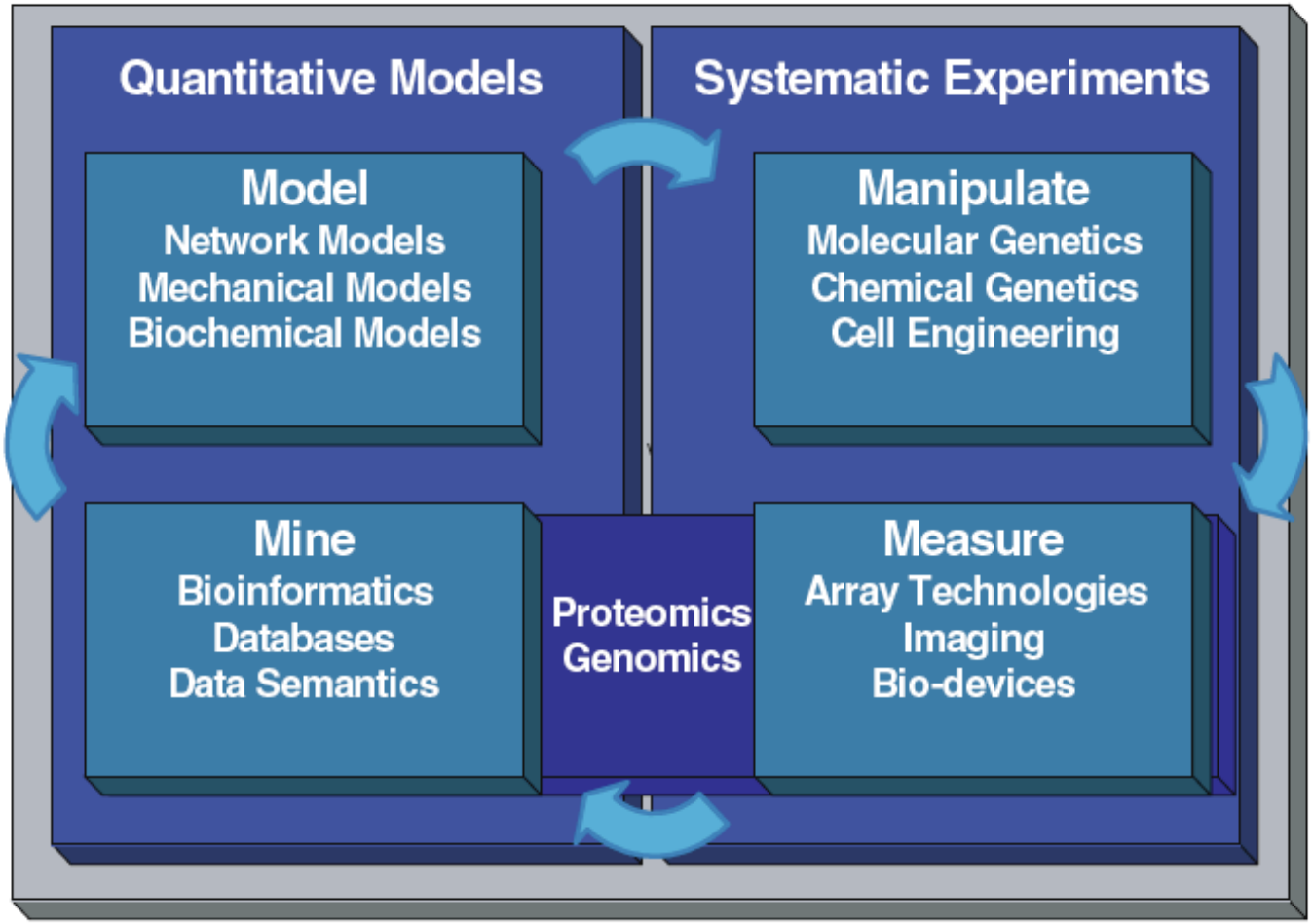
6. Golgi Apparatus (packaging center)

7. Vesicles (transportation of products)

8. Cell Membrane (security gate)

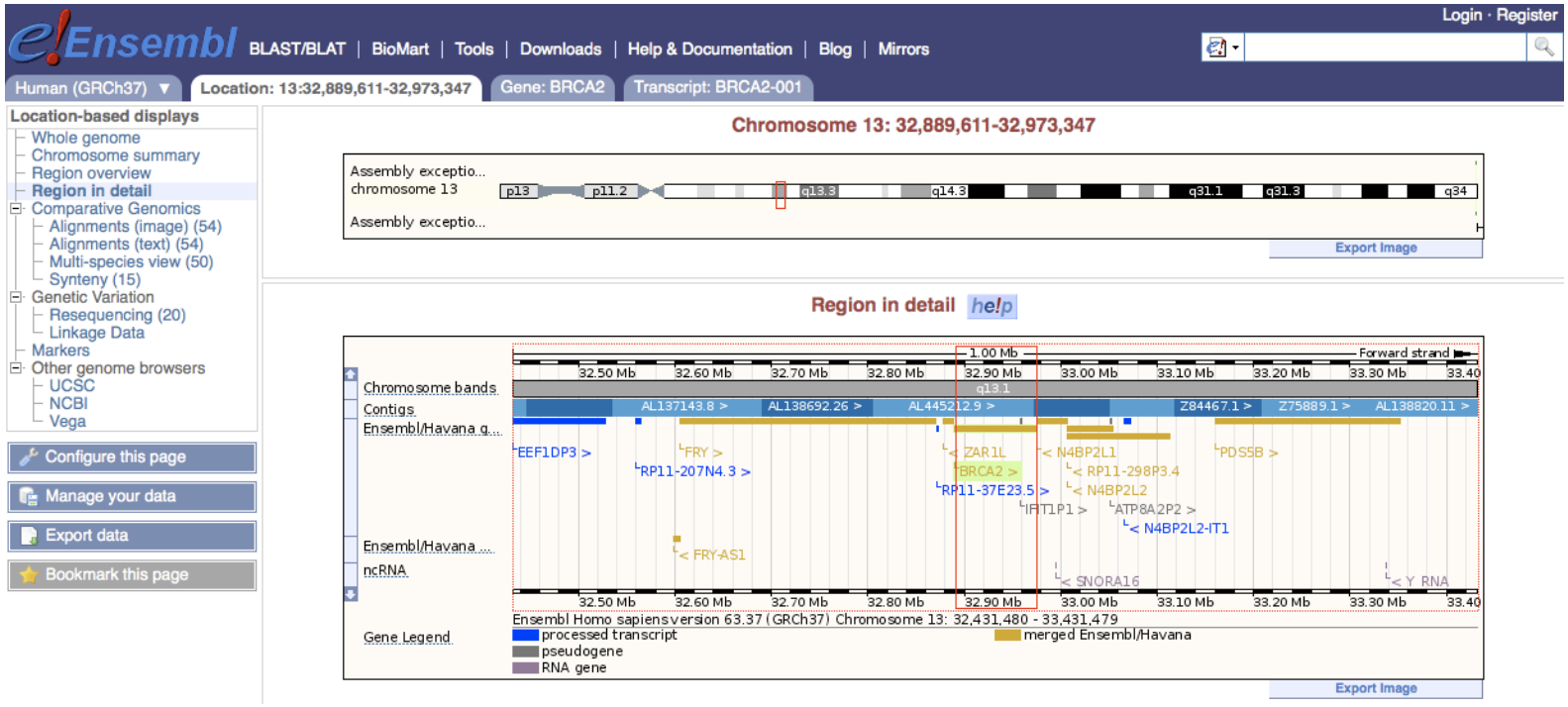
https://tsiantific.wordpress.com/tag/cells/

4Ms: operational definition



Ideker *et al. Annals of Biomedical Engineering (*2006) 34(2):257-64

Typical Workflow of Bioinformatics



- Organize / store / retrieve biological information from databases and public domains
- Perform some basic statistical testing

“Synthetic Biology”

- A discipline that uses engineering principles to design and assemble improved biological components and systems

- Artificially design and construction of biological devices systems, and machines

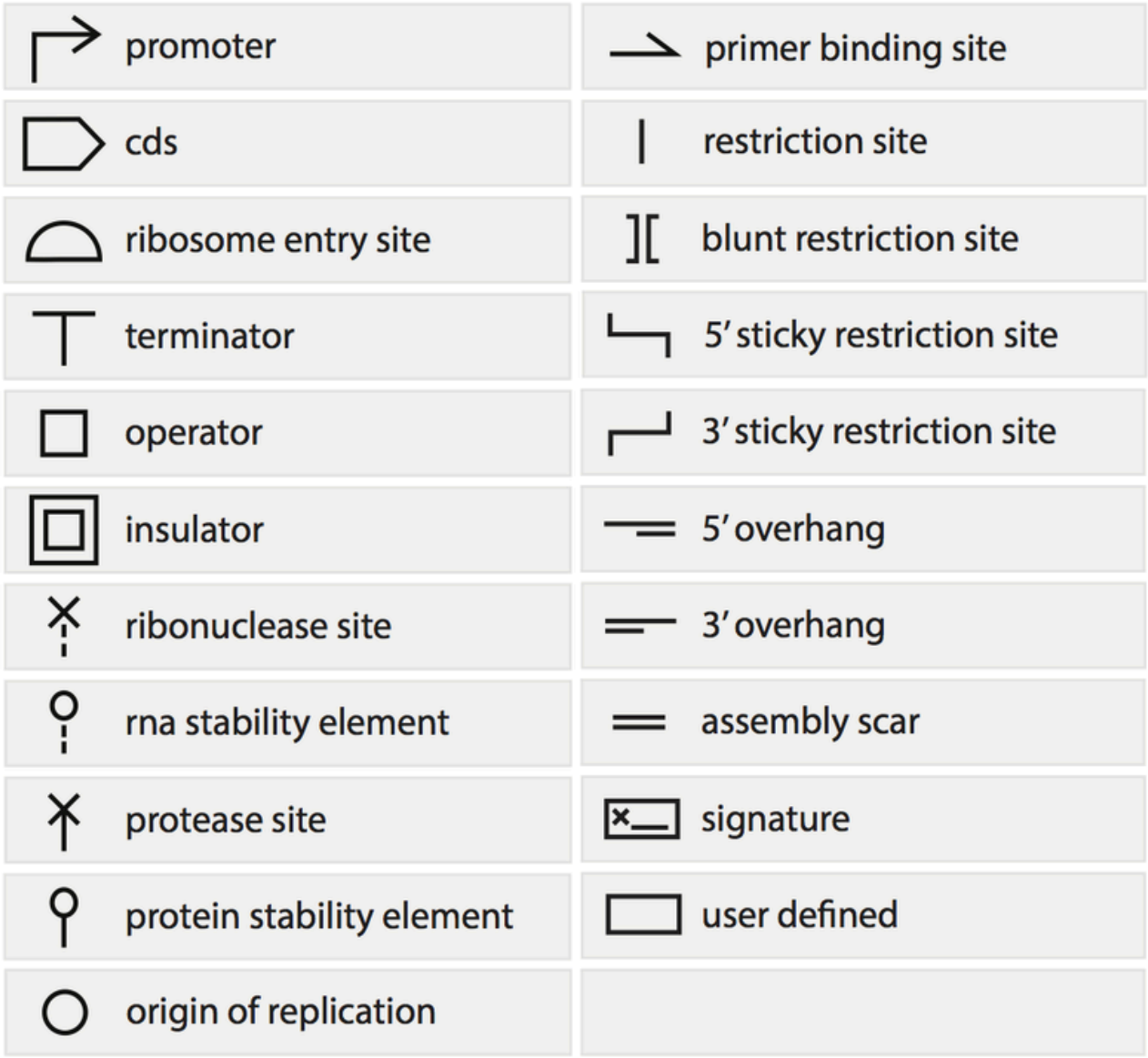
- Combining interdisciplinary methods from biotechnology, genetic engineering, molecular biology, biophysics

Biobricks

- Standardized biological parts are DNA sequences

- Common interface and are designed to be composed and incorporated into living cells

Biobricks



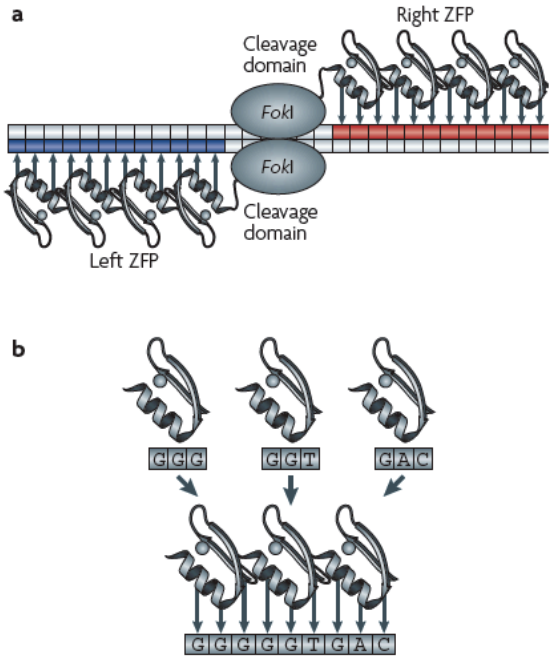
Synthetic Biology Open Language (SBOL)

Genome Editing and Reverse Engineering

- Zinc finger nucleases (ZFNs)

- Transcription Activator-Like Effector Nucleases (TALENs)

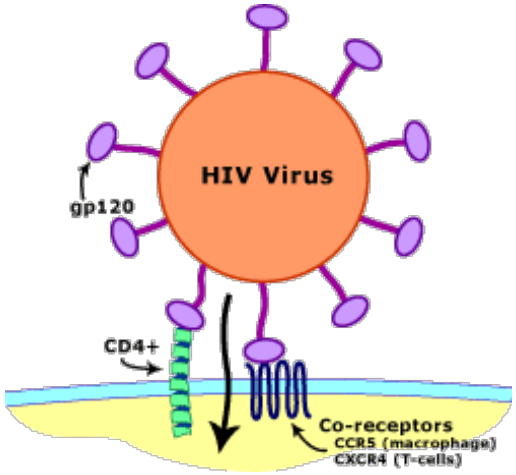
- CRISPR/Cas system



nature biotechnology

Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases

Elena E Perez[1,2], Jianbin Wang[3], Jeffrey C Miller[3], Yann Jouvenot[3,4], Kenneth A Kim[3], Olga Liu[1], Nathaniel Wang[3], Gary Lee[3], Victor V Bartsevich[3], Ya-Li Lee[3], Dmitry Y Guschin[3], Igor Rupniewski[3], Adam J Waite[3], Carmine Carpenito[1], Richard G Carroll[1], Jordan S Orange[2], Fyodor D Urnov[3], Edward J Rebar[3], Dale Ando[3], Philip D Gregory[3], James L Riley[1], Michael C Holmes[3] & Carl H June[1]



Pict from: walmartramen.blogspot.com/2014/03/bone-marrow-transplant-and-engineered.html

Rewriting Life

CRISPR in 2018: Coming to a Human Near You

The first clinical trials are slated to begin in the U.S. and Europe while others are stalled.

by Emily Mullin December 18, 2017

https://www.technologyreview.com/s/609722/crispr-in-2018-coming-to-a-human-near-you/

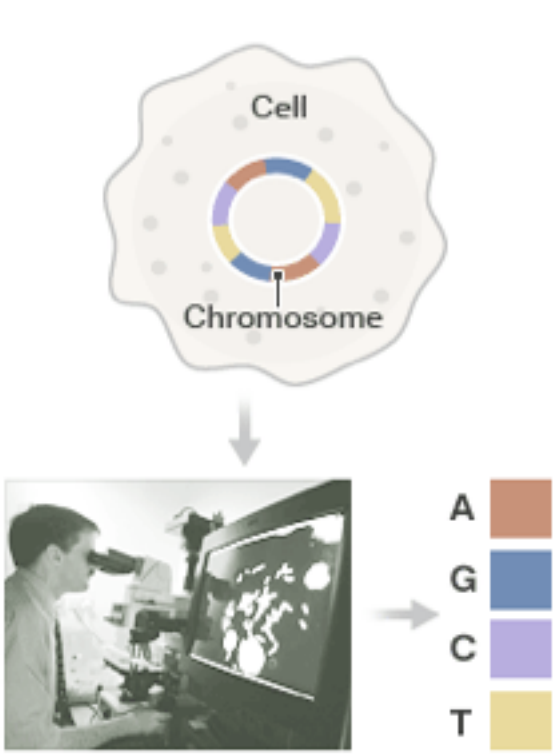
BETTER FASTER CRISPR

Chinese scientists used Crispr gene editing on 86 human patients

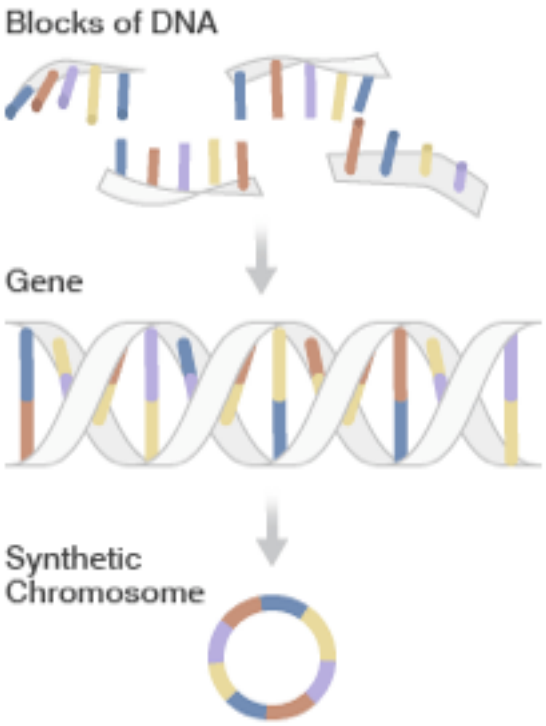
By Katherine Ellen Foley · January 23, 2018

https://qz.com/1185488/chinese-scientists-used-crispr-gene-editing-on-86-human-patients/

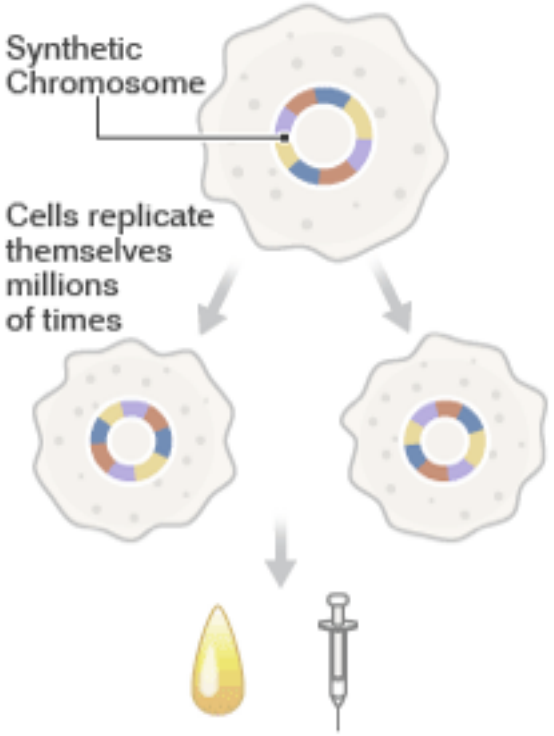
Venter's artificial life?



The scientists "decoded" the chromosome of an existing bacterial cell - using a computer to read each of the letters of genetic code.



They copied this code and chemically constructed a new synthetic chromosome, piecing together blocks of DNA.

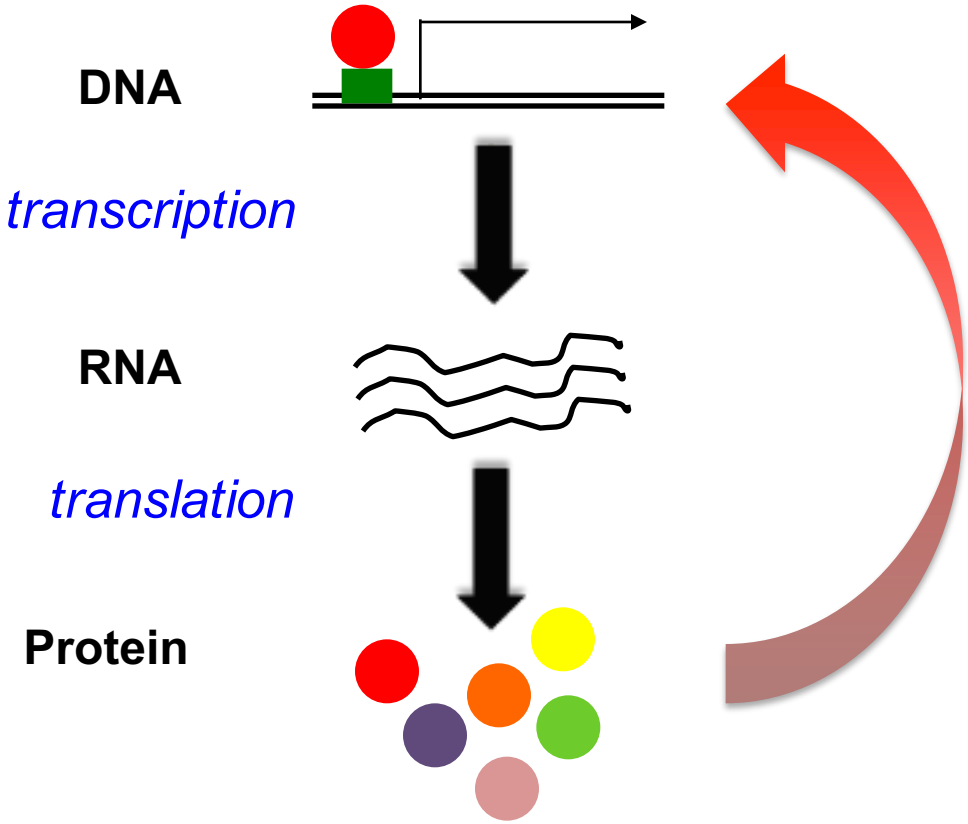


The team inserted this chromosome into a bacterial cell which replicated itself. Synthetic bacteria might be used to make new fuels and drugs.

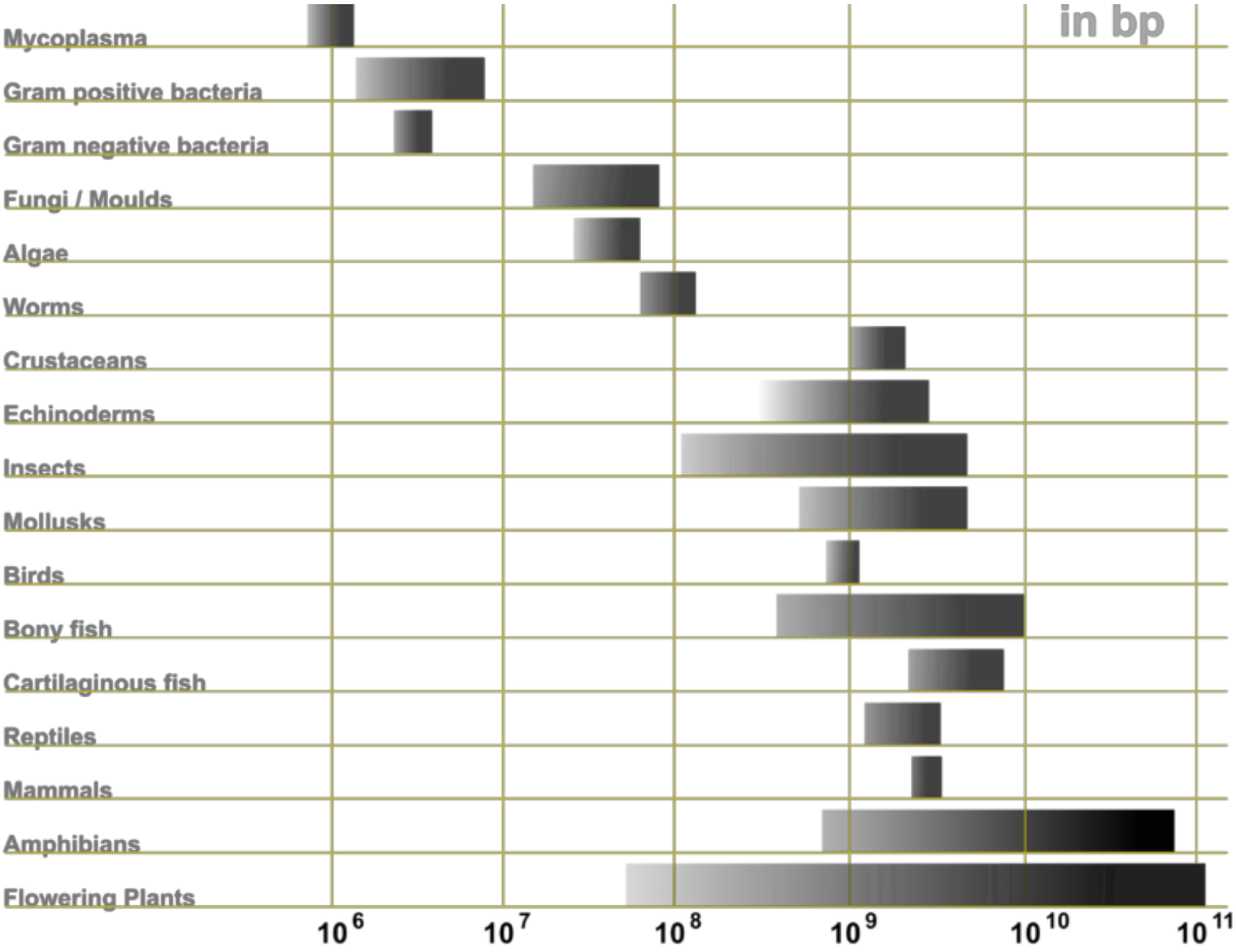
Varodom Charoensawan, Mahidol University

Illustrations from www.bbc.co.uk

The world of “Omics” data



Genome size (bp)



http://en.wikipedia.org/wiki/File:Genome_Sizes.png

TABLE 1–2 A Few of the Many Organisms Whose Genomes Have Been Completely Sequenced

Organism	Genome size (millions of nucleotide pairs)	Number of genes	Biological interest
Mycoplasma genitalium	0.58	483	Smallest true organism
Treponema pallidum	1.1	1,039	Causes syphilis
Borrelia burgdorferi	1.44	1,738	Causes Lyme disease
Helicobacter pylori	1.7	1,589	Causes gastric ulcers
Methanococcus jannaschii	1.7	1,783	Archaean; grows at 85 °C!
Haemophilus influenzae	1.8	1,738	Causes bacterial influenza
*Archaeoglobus fulgidus**	2.2	—	High-temperature methanogen
Synechocystis sp.	3.6	4,003	Cyanobacterium
Bacillus subtilis	4.2	4,779	Common soil bacterium
Escherichia coli	4.6	4,377	Some strains cause toxic shock syndrome
Saccharomyces cerevisiae	12.5	5,770	Unicellular eukaryote
Plasmodium falciparum	23	5,268	Causes human malaria
Caenorhabditis elegans	100	19,400	Multicellular roundworm
Anopheles gambiae	278	13,700	Malaria vector
Arabidopsis thaliana	157	25,500	Model plant
Oryza sativa	390	37,500	Rice
Drosophila melanogaster	140	13,000	Laboratory fly ("fruit fly")
Mus musculus domesticus	2.4×10^3	25,000	Laboratory mouse
Pan troglodytes	2.4×10^3	25,000	Chimpanzee
Homo sapiens	2.9×10^3	25,000	Human

*The number of genes is not yet determined.

Table 1-2

Lehninger Principles of Biochemistry, Fifth Edition

High-throughput sequencing technologies (aka Next-generation sequencing, NGS)

illumina®

PACIFIC BIOSCIENCES®

www.technologyreview.com
www.gatc-biotech.com/

Illumina sequencing

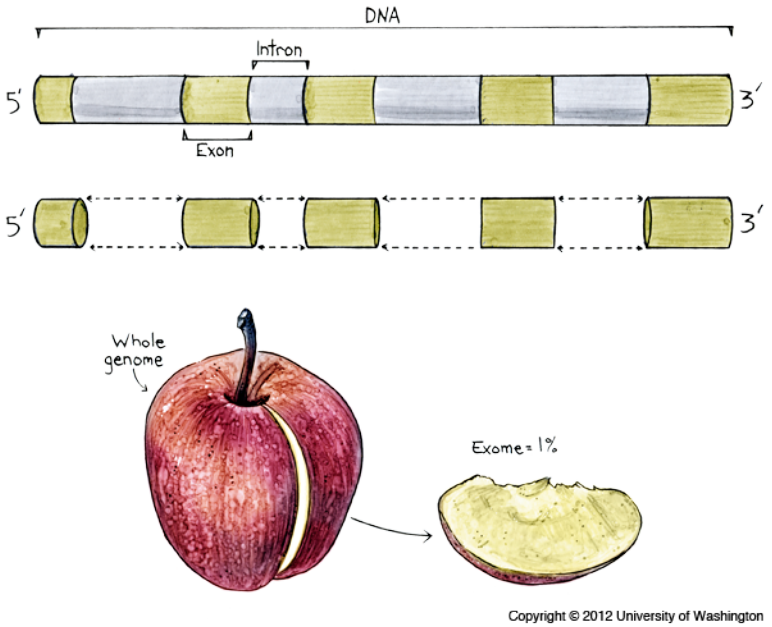
1.5-2 TB of raw data per run (3 days)

illumina®

Exome sequencing

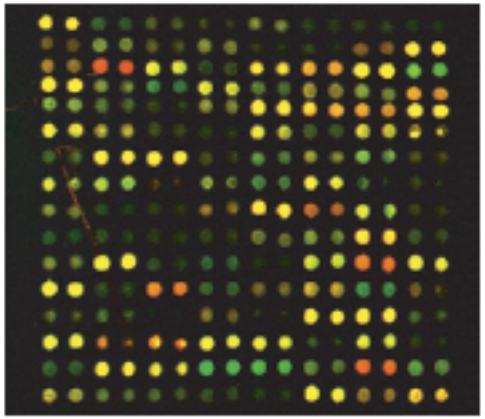
- Key: Enriched for protein-coding regions of the genome
- Humans have about 180,000 exons
- That is, about 1% of the human genome

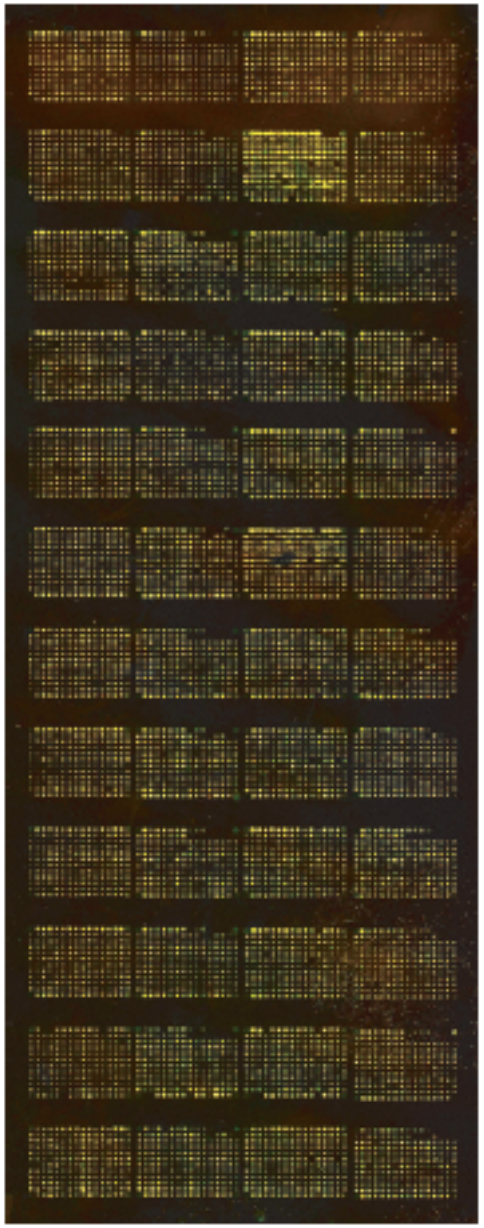
(30 million base pairs)



https://www.my46.org/intro/whole-genome-and-exome-sequencing

Microarray





Gene expression, genome variation, etc.

Typical data per chip: 10 - 100 MB

Pictures from Stekel, Microarray bioinformatics, 2003.

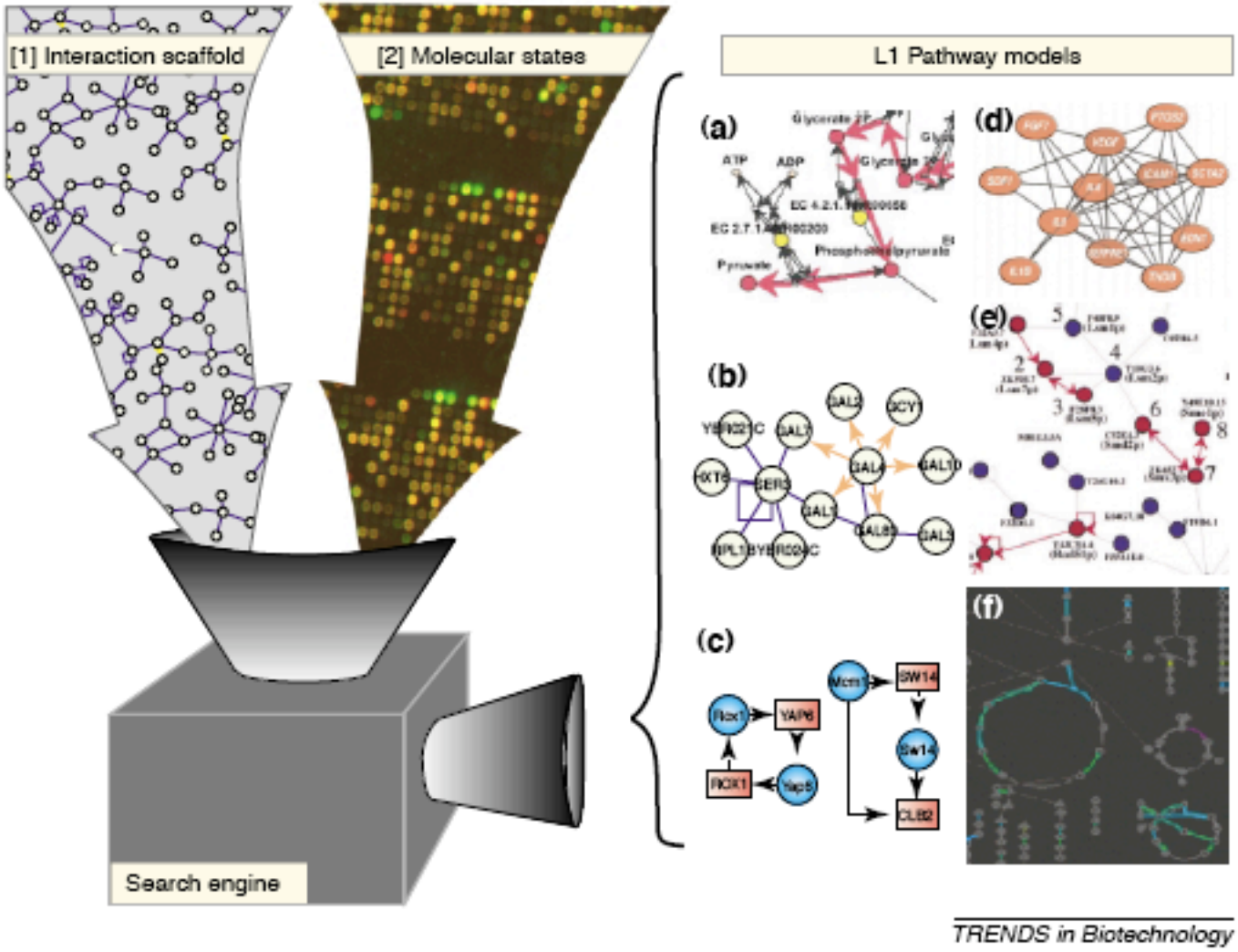
High-throughput proteomics

LC-MS/MS

LTQ Orbitrap

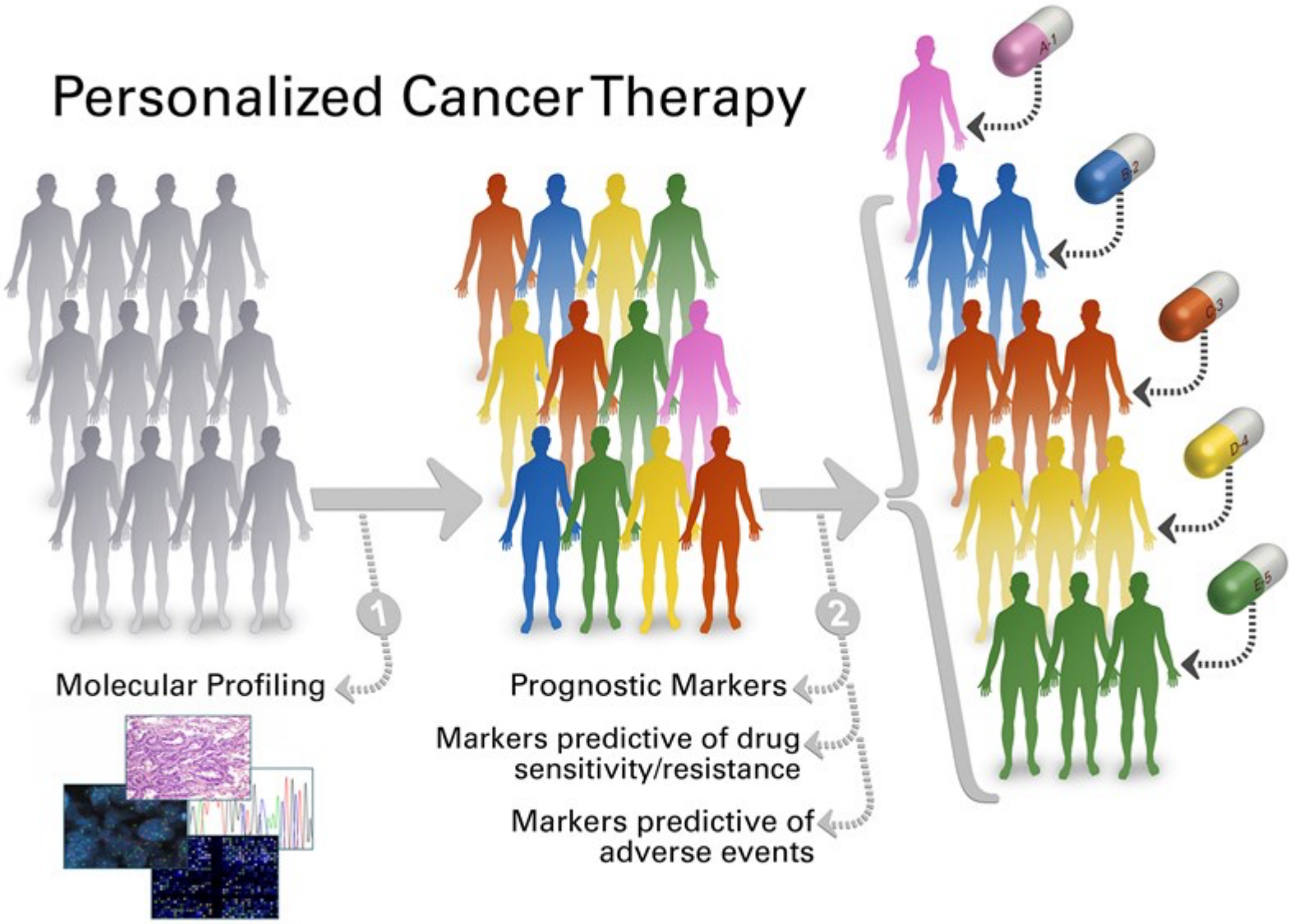
http://www.thermoscientific.com/en/products/mass-spectrometry.html

From large-scale data to meaningful findings



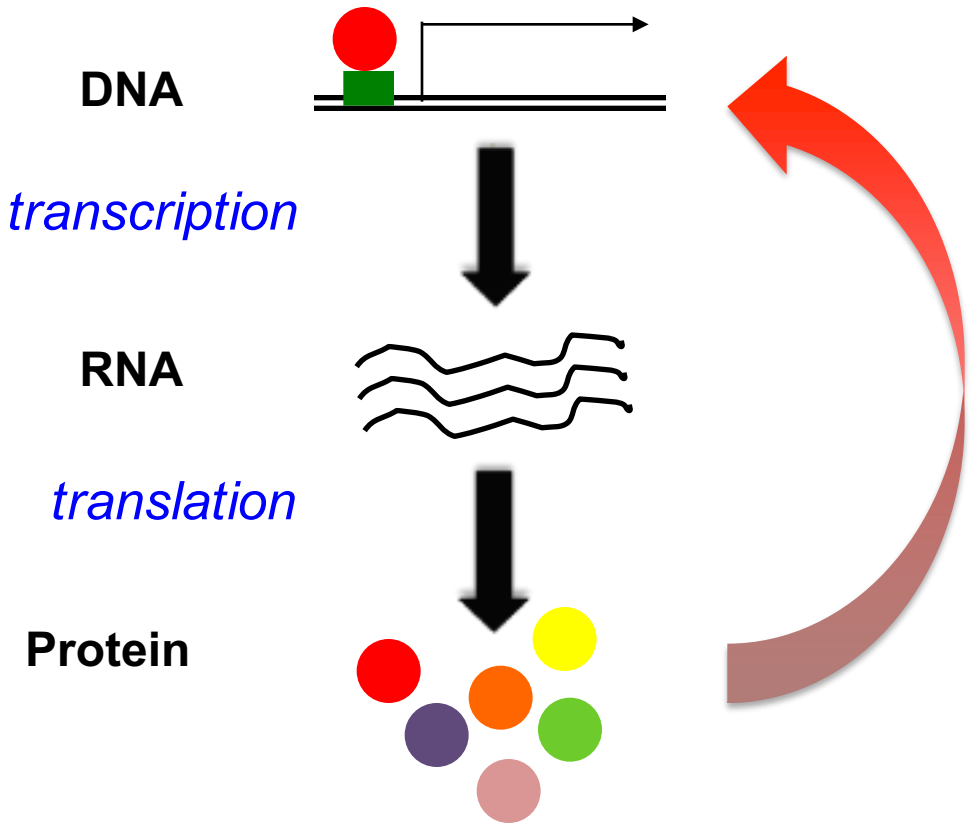
Ideker *et al. Trends in Biotech (*2003) Jun;21(6):255-62.

Personalized medicine?

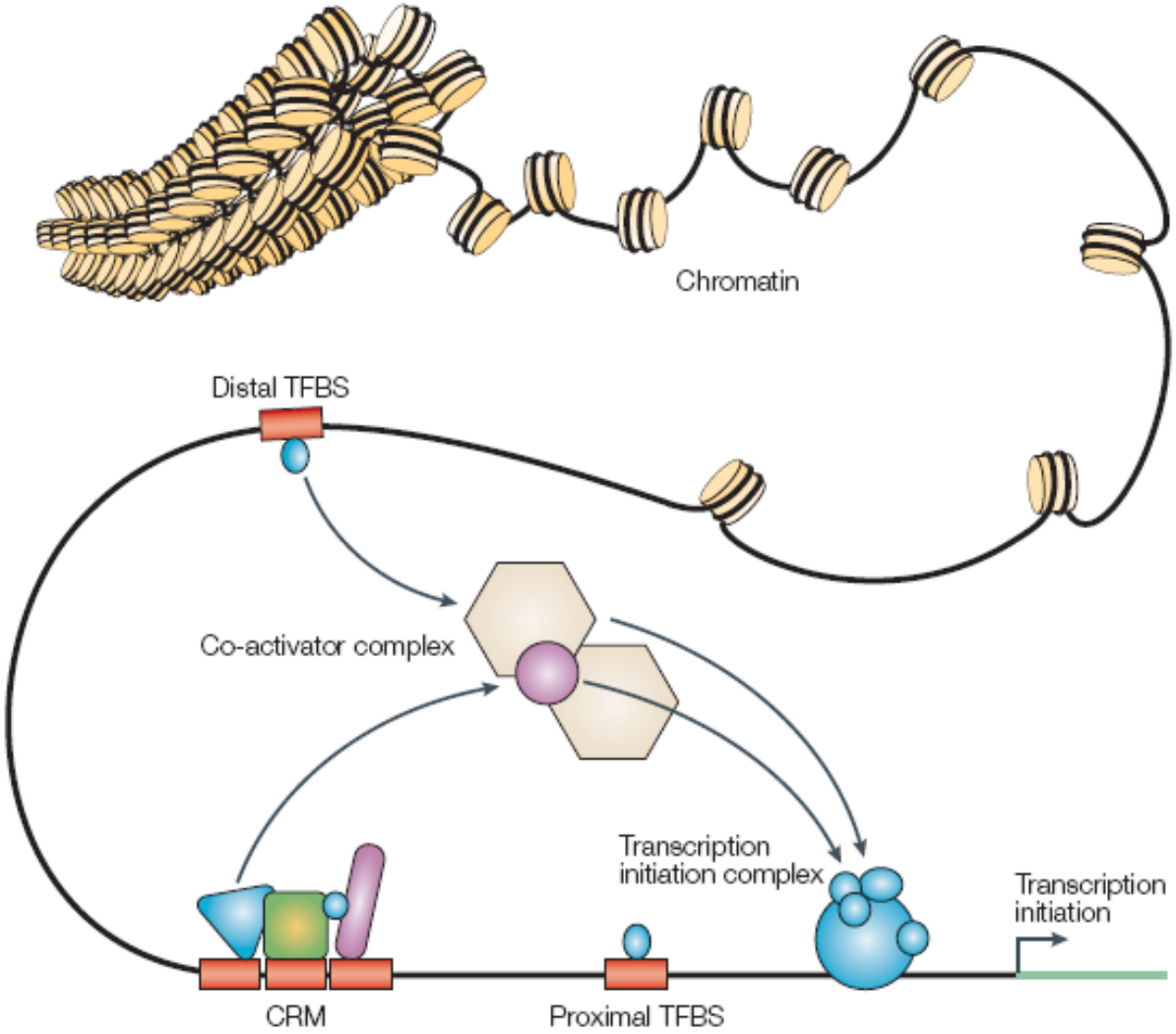


Pict from: gahtour.com/iran-looks-personalized-cancer-medicine-major-treatment/

Central dogma of molecular biology

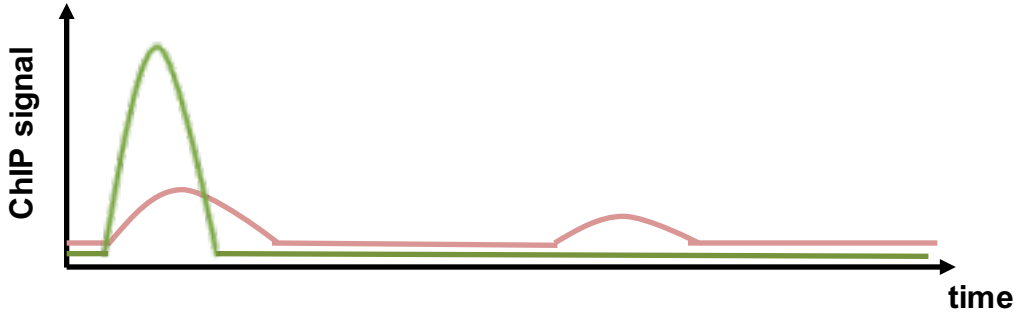


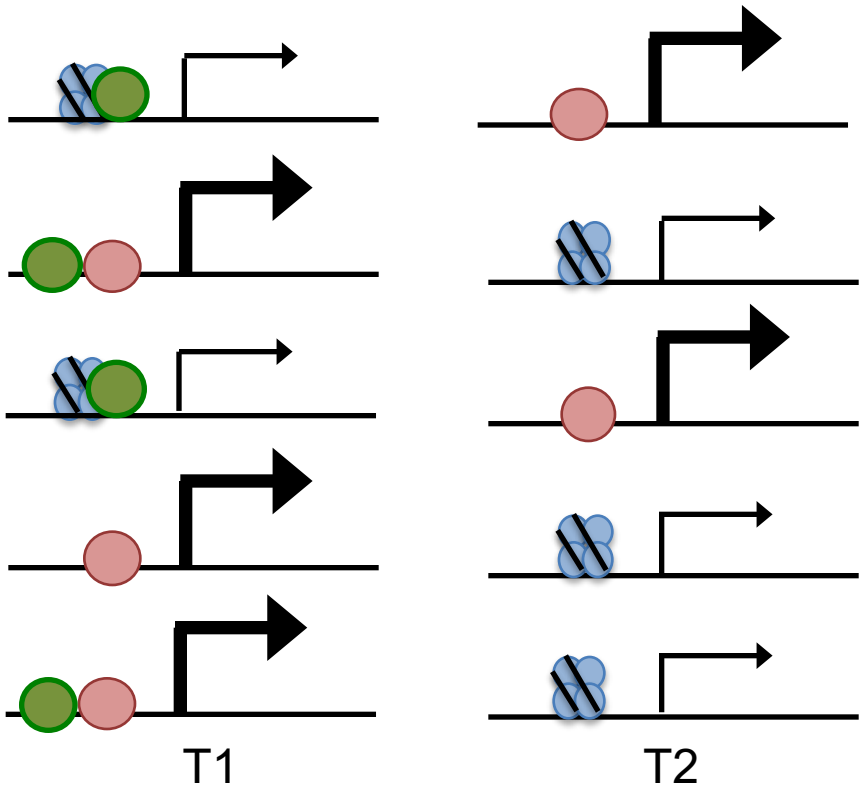
Transcription Factors Play a Central Role in Differential Expression of Genes



Wasserman and Sandelin (2004) *Nat Rev Genet.* **5***:* 276—287

Transcriptional Regulation is highly dynamics





T1-T2 = seconds

Charoensawan *et al. Bioessays* (2015)

Leading Edge
Review

Cell

Transcriptional Regulation and Its Misregulation in Disease

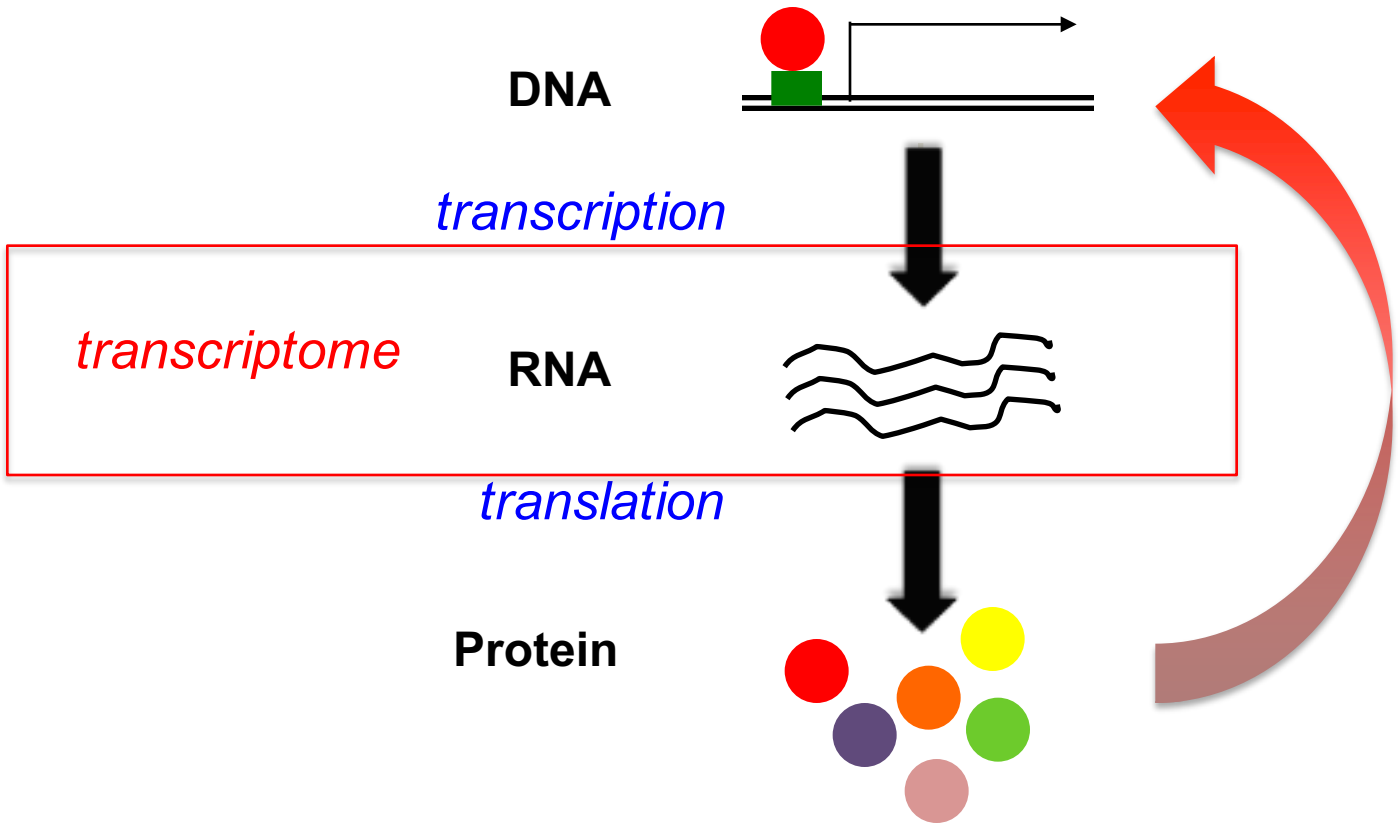
Tong Ihn Lee[1] and Richard A. Young[1,2,*]

- **Various types of Cancers:**
  - NF-kappaB
  - STAT family proteins
  - P53
  - Steroid receptors
- **Autoimmunity and Inflammation**
- **Developmental Disorders**
- **Diabetes**
- **Cardiovascular Diseases**

Outlines

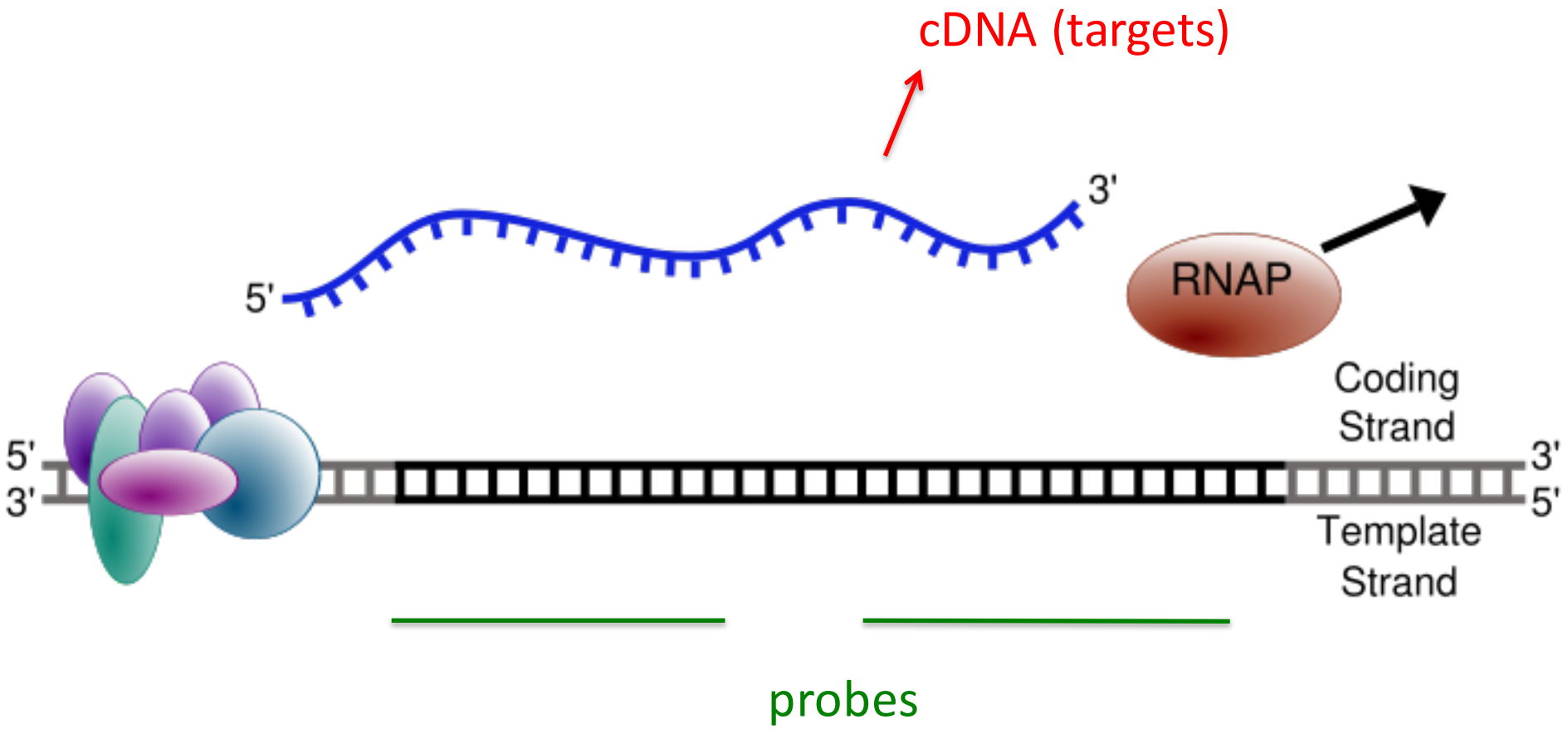
- Transcriptomics: genome-wide transcriptional read-outs
- Regulation of transcription (and transcriptomes)
- Putting things in perspectives

Central dogma of molecular biology



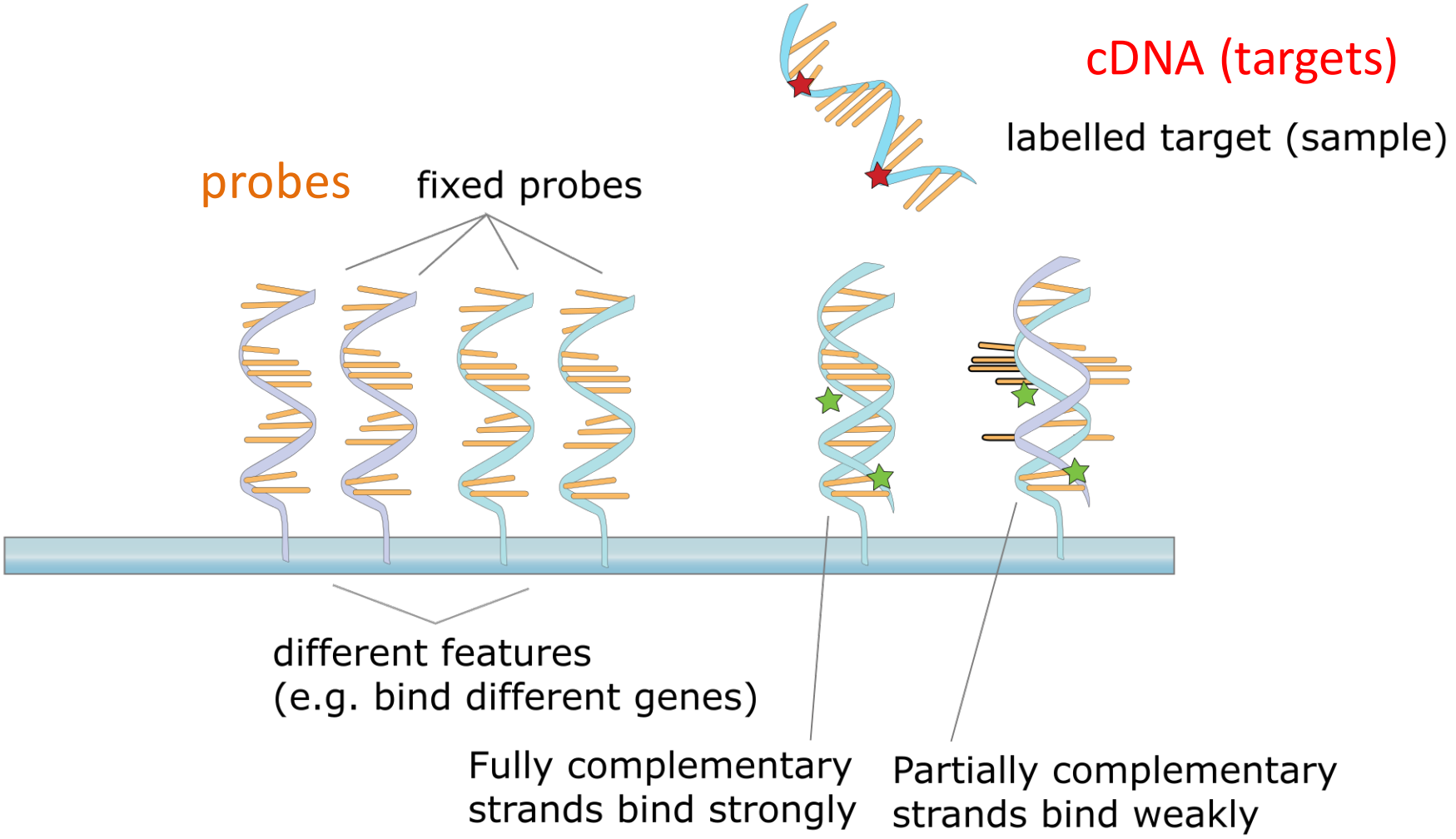
How do we determine transcriptional rates?

Expression Array / Gene Array

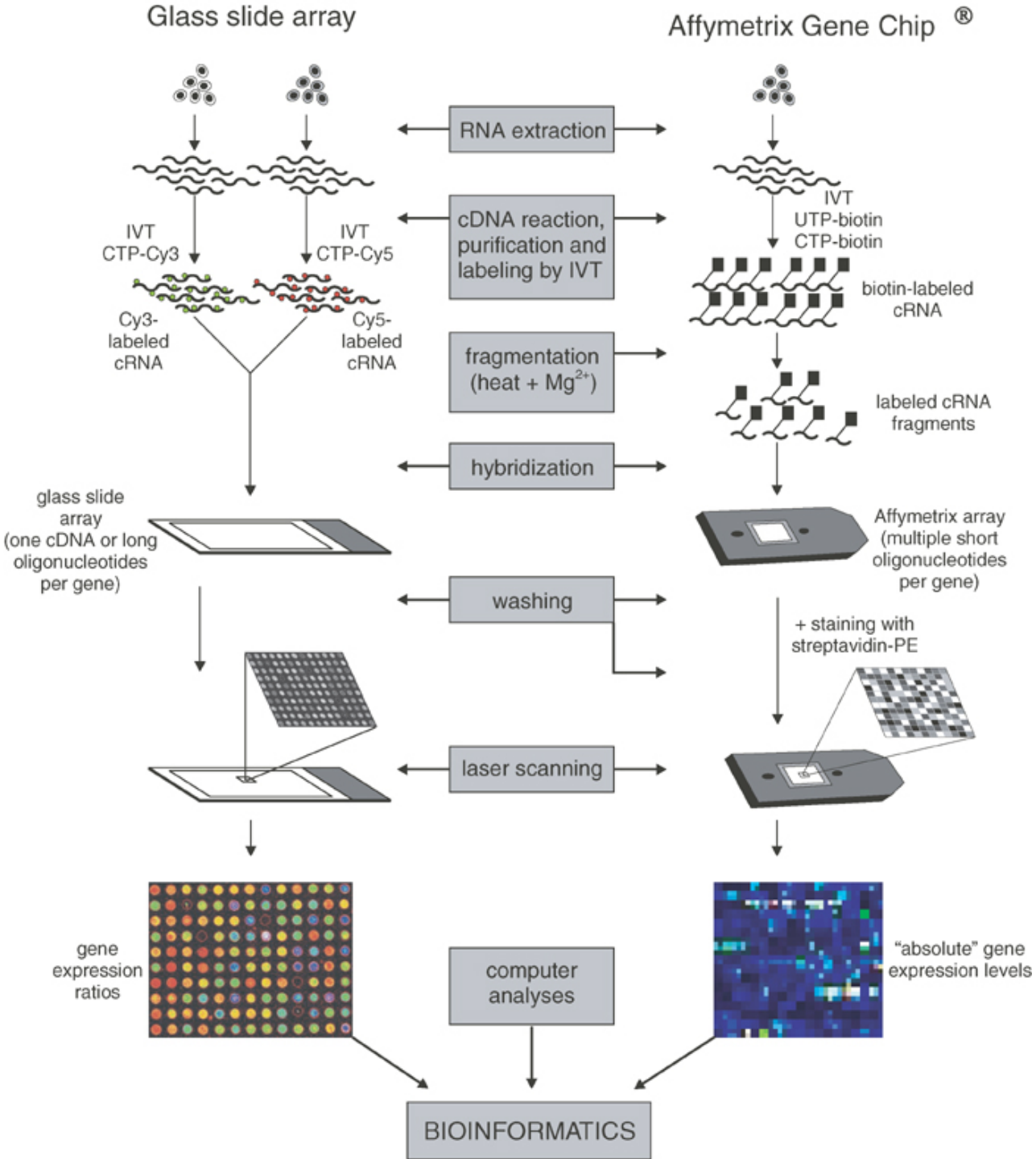


Modified from Forluvoft 2007 via Wikimedia commons

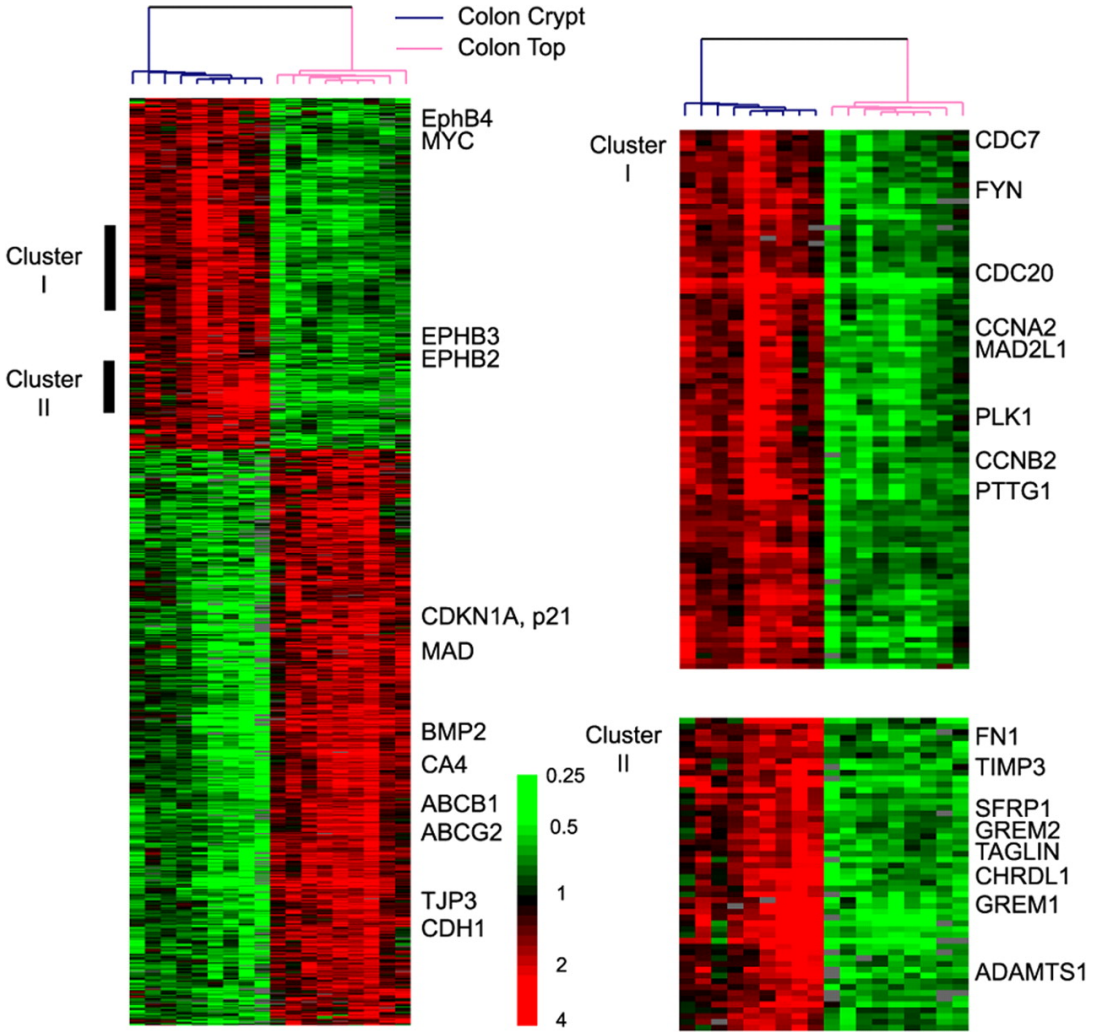
Microarrav in Nutshell



Modified from Squidonius 2009 via Wikimedia commons

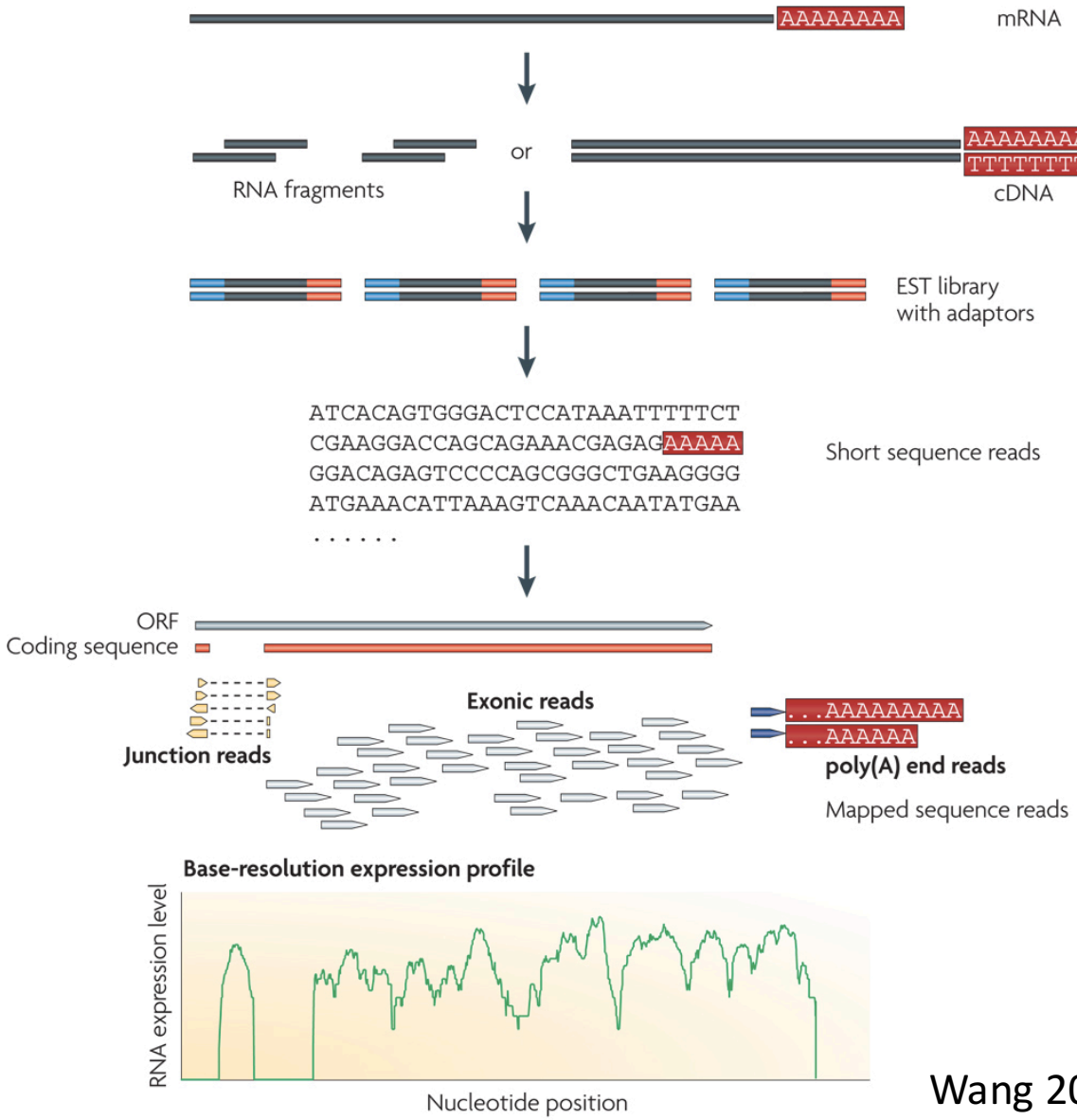


Discovery of new Biomarkers from Expression Array



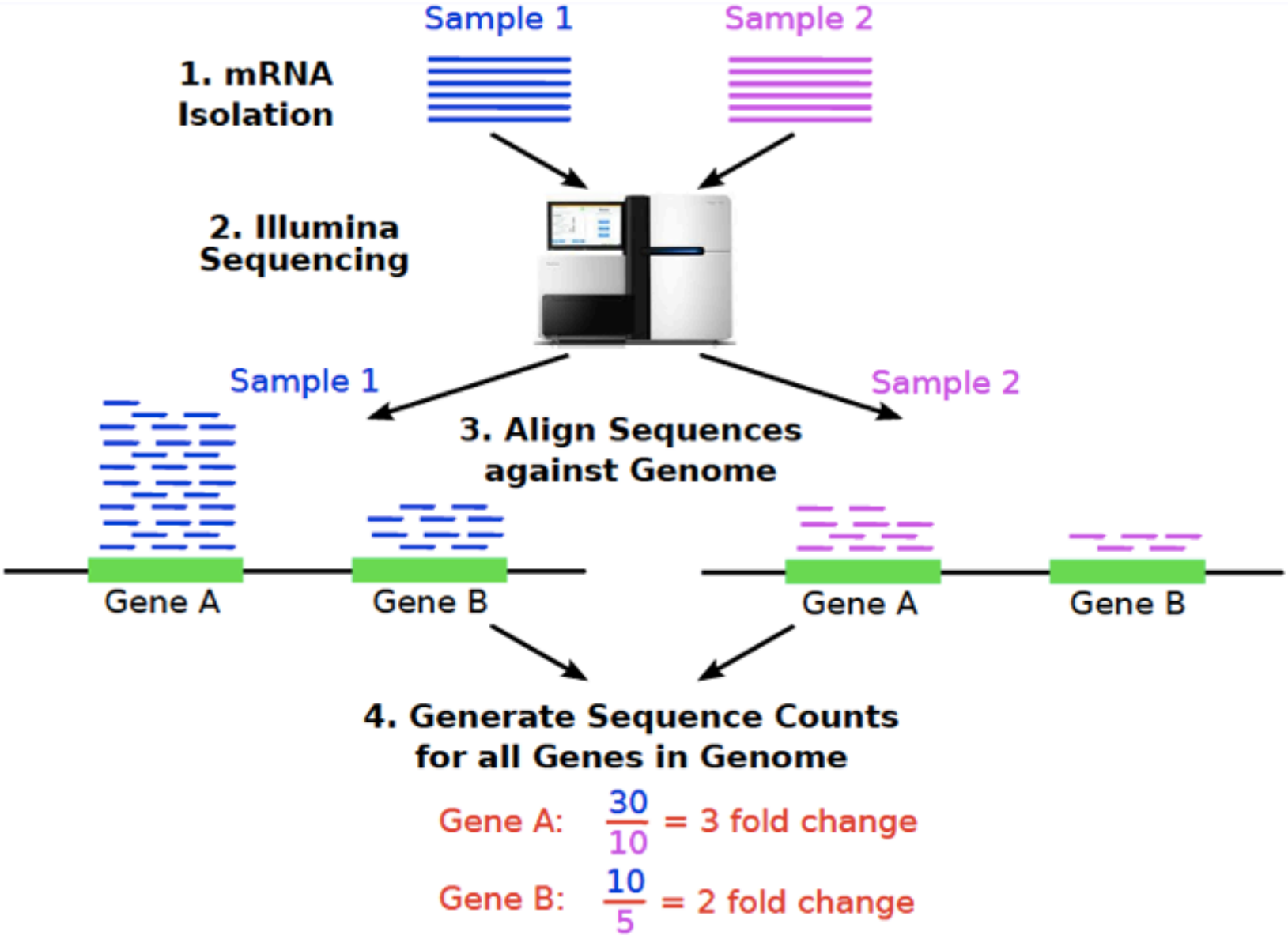
Cynthia Kosinski et al. PNAS 2007;104:15418-15423

RNA-seq: whole transcriptome shotgun sequencing



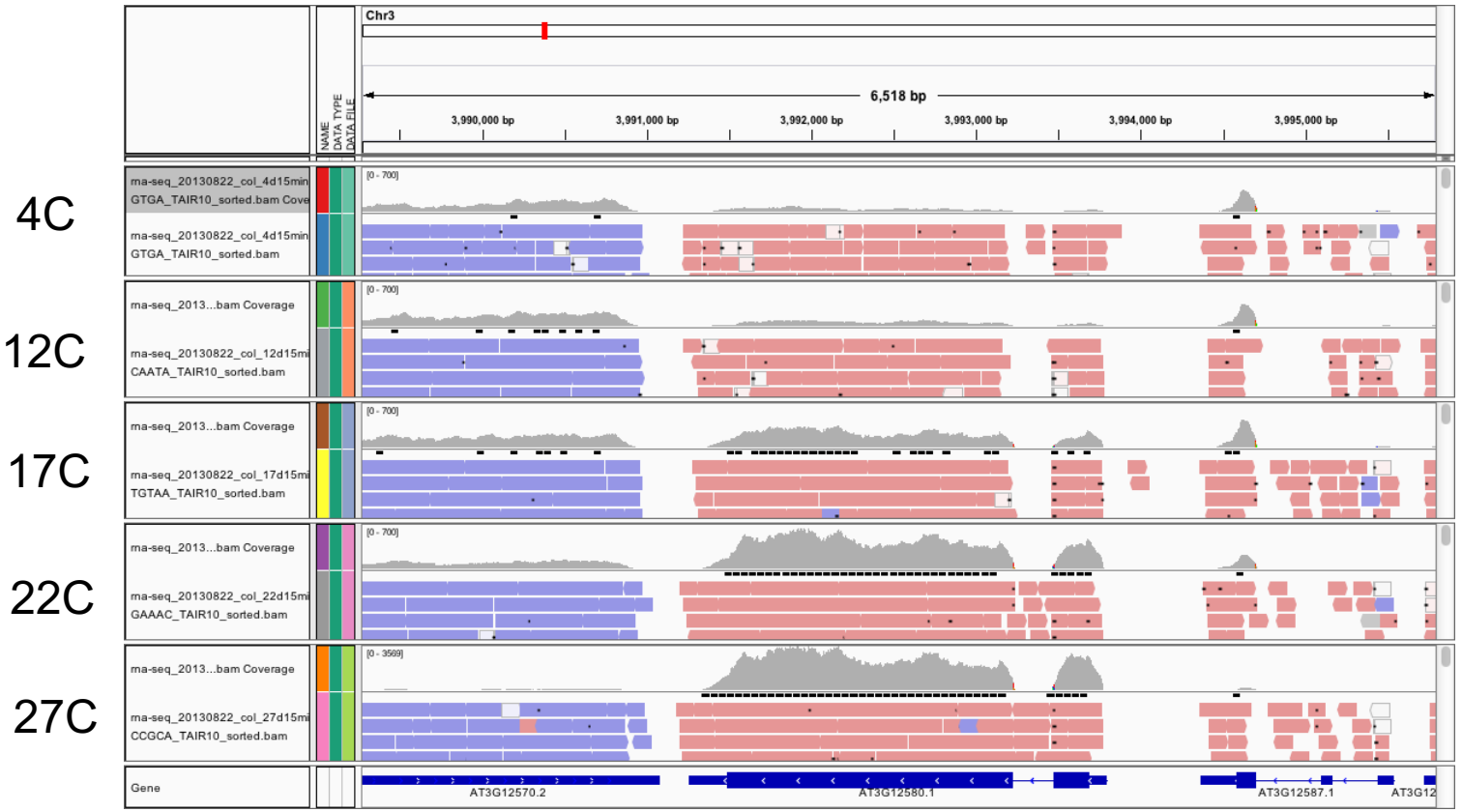
Wang 2009 *Nat Rev Genet.*

Differential Expressed mRNA



Thomas Girke, Analysis of RNA-Seq Data with R/Bioconductor, December 14, 2013)

Examples of RNA-seq data: HSP70 as a temperature sensor



Transcriptional response after 15 mins

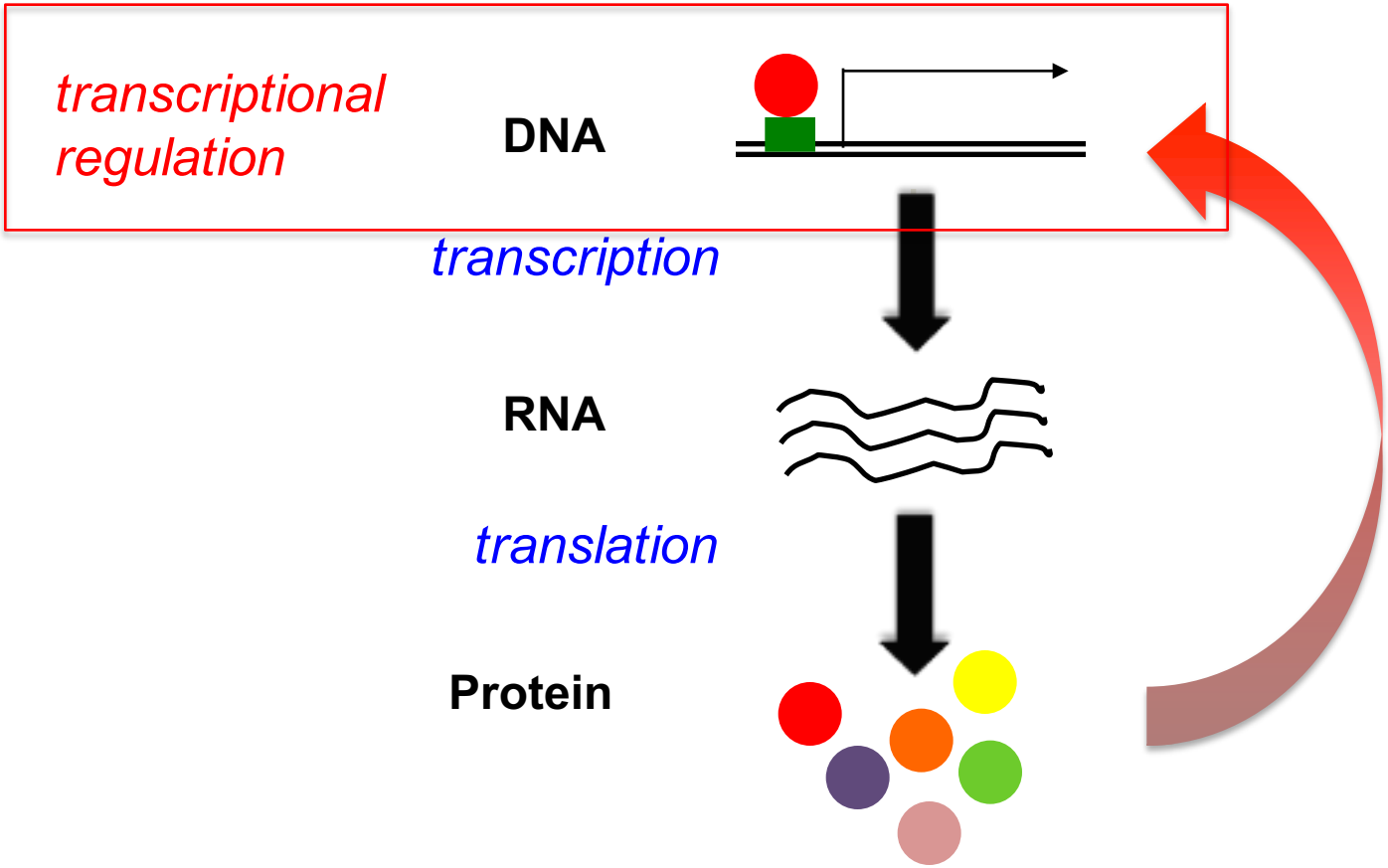
(Scale 0-600 for all, except 27C: 0-3500)

With Sandra Cortijo and Phil Wigge

Outlines

- Transcriptomics: genome-wide transcriptional read-outs (microarray, RNA-seq)
- Regulation of transcription (and transcriptomes)
- Putting things in perspectives

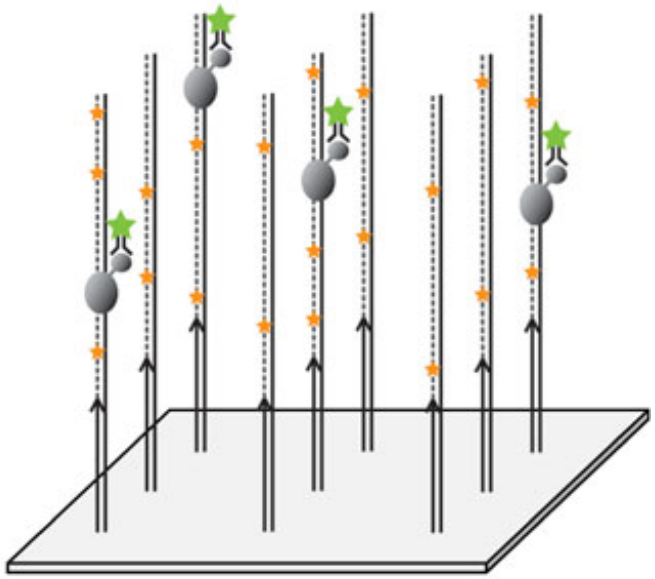
Central dogma of molecular biology



How to determine protein-DNA interaction?

in vitro

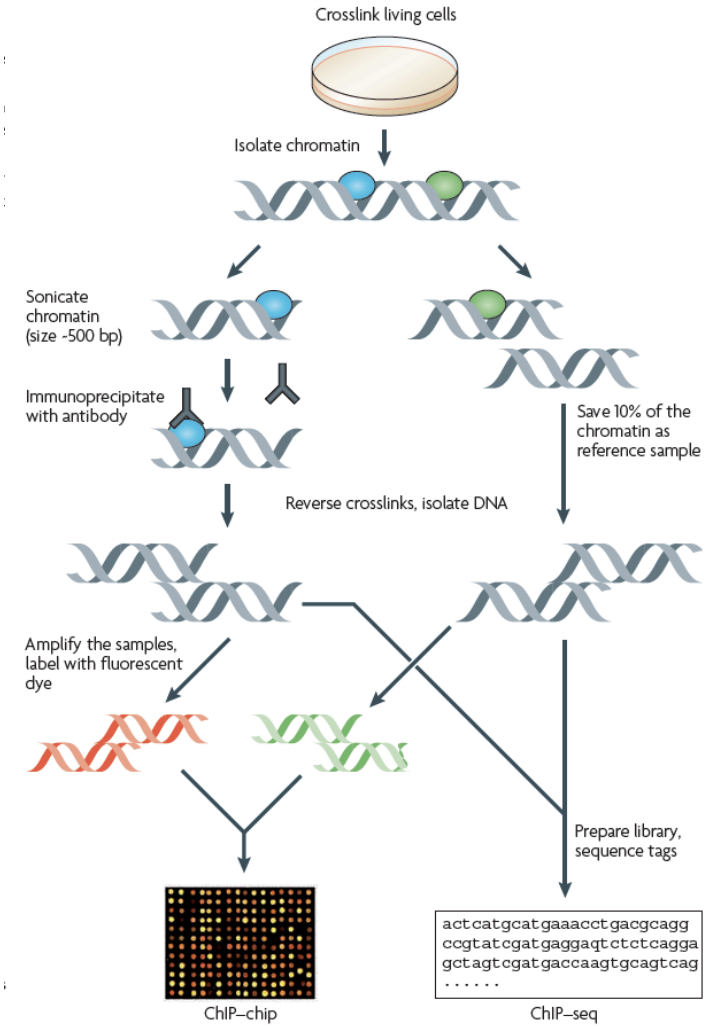
purified TF / naked DNA



e.g. Protein Binding Microarray, EMSA

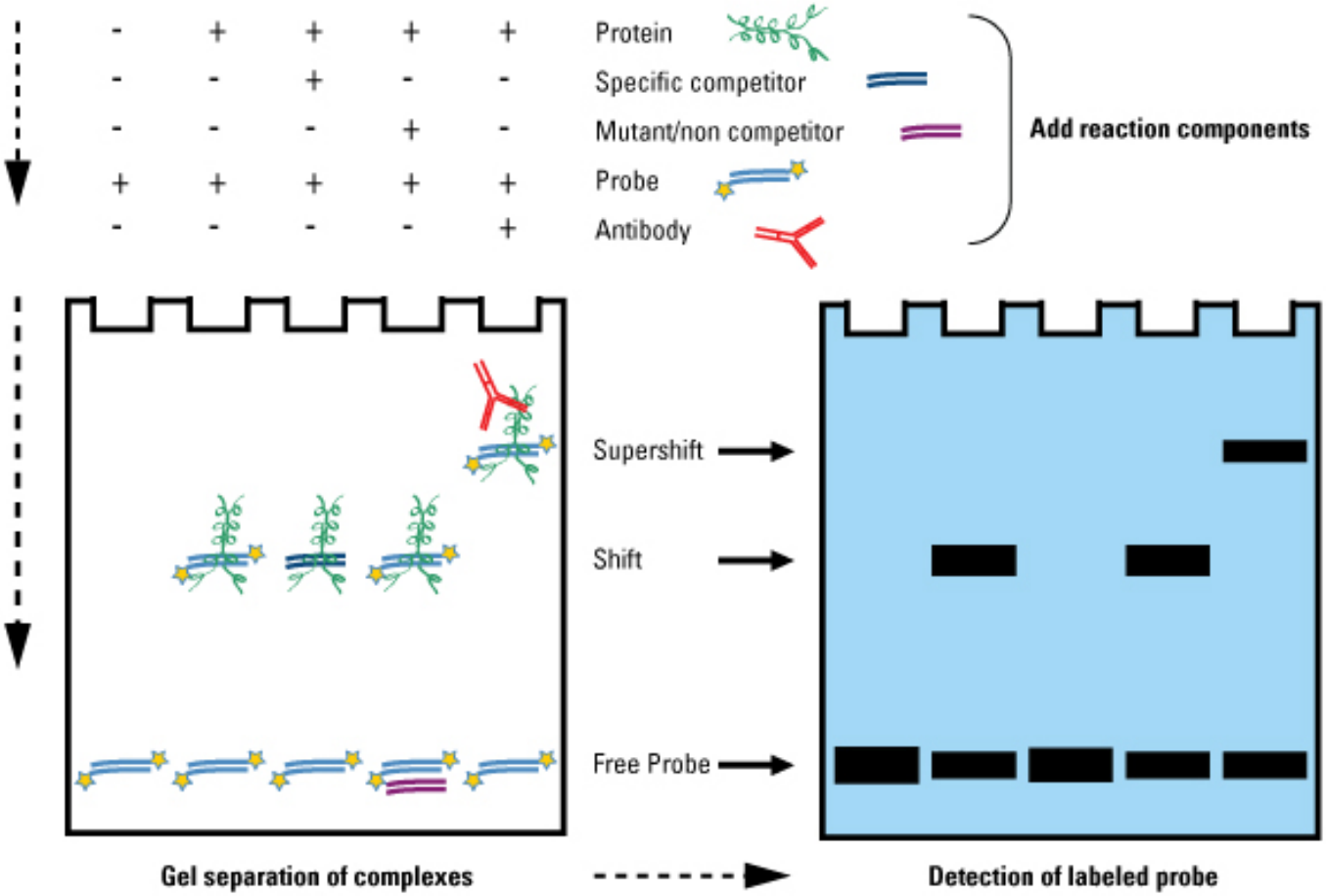
in vivo

all TFs / chromatin



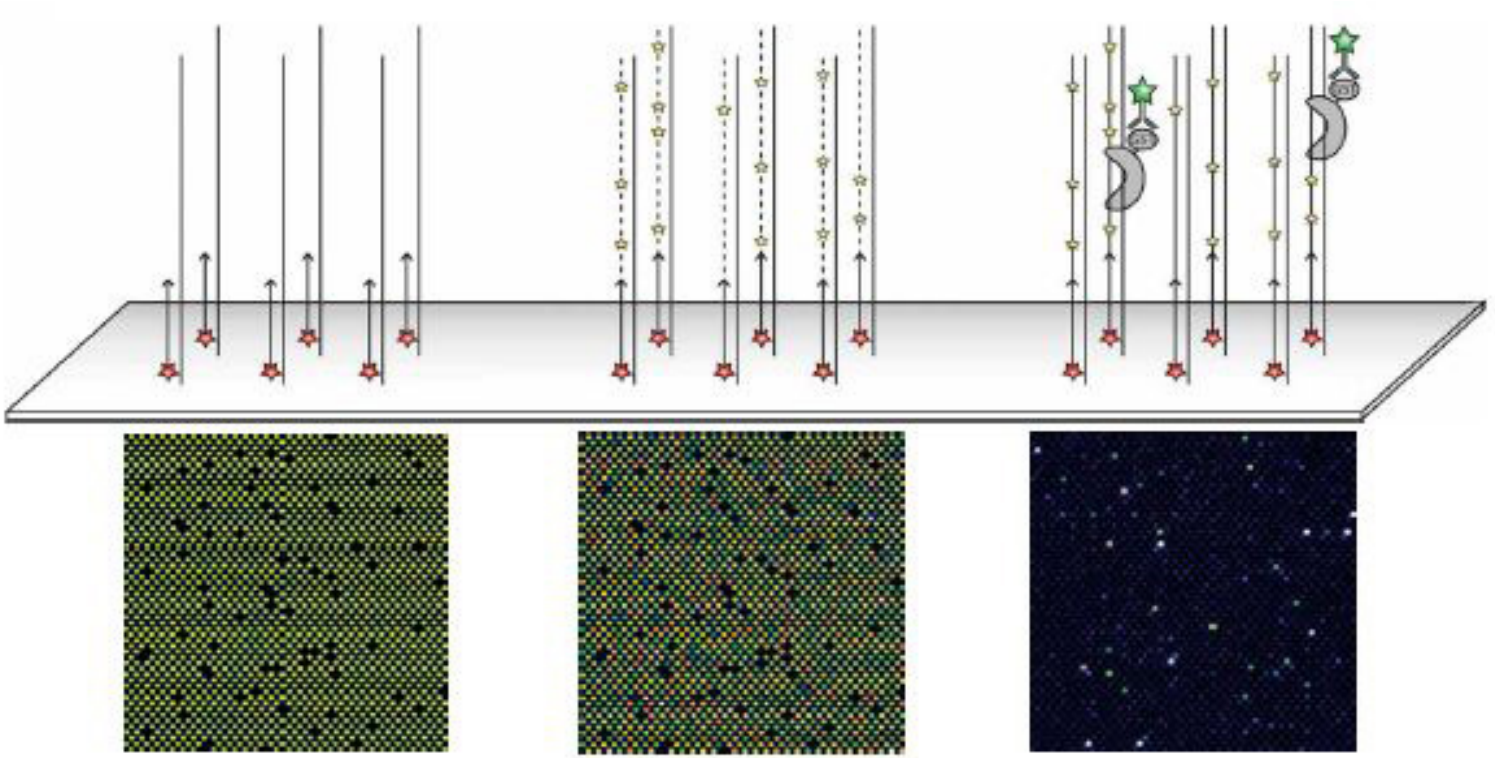
e.g. ChIP-chip, ChIP-seq

Electrophoretic mobility shift assay (EMSA)



https://www.thermofisher.com/th/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/

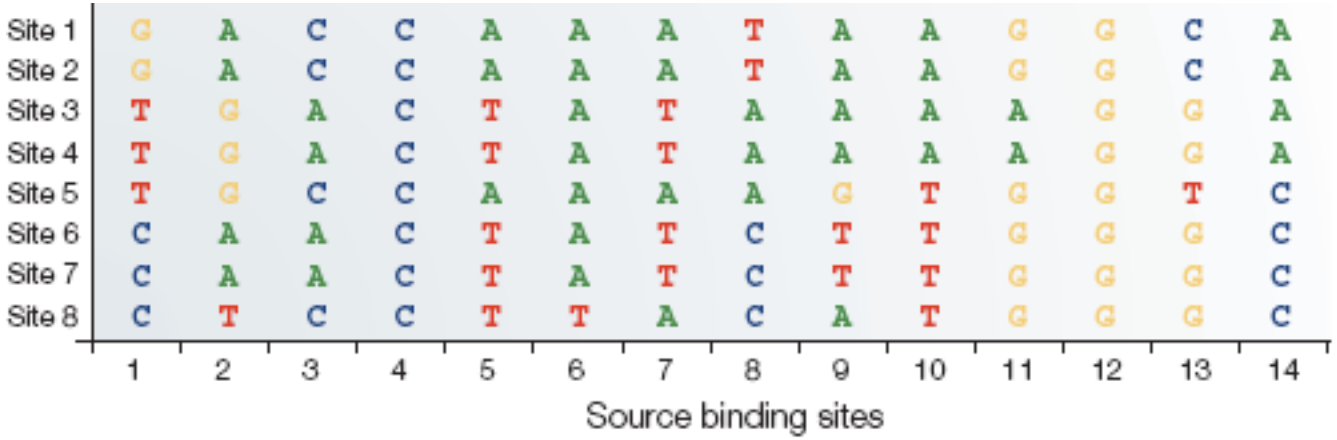
Protein Binding Microarray (PBM)



http://the_brain.bwh.harvard.edu/

TF-DNA binding specificity (1)

Alignment of TF-bound sequences



Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

Position weight matrix (PWM)

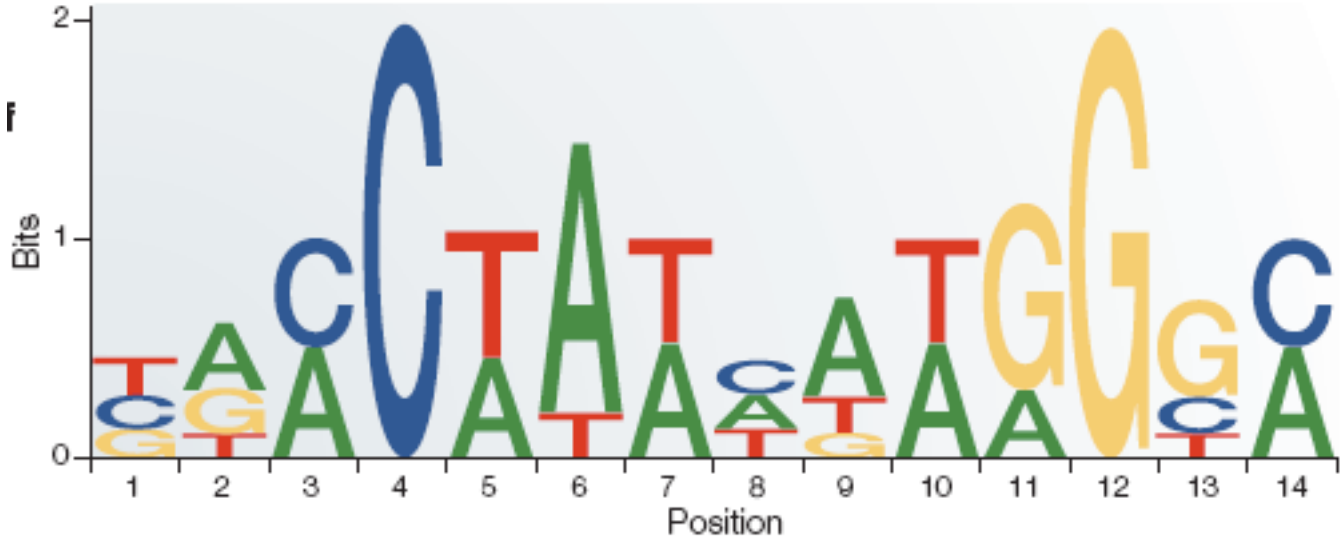
A	−1.93	0.79	0.79	−1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	−1.93	−1.93	0.79
C	0.45	−1.93	0.79	1.68	−1.93	−1.93	−1.93	0.45	−1.93	−1.93	−1.93	−1.93	0.00	0.79
G	0.00	0.45	−1.93	−1.93	−1.93	−1.93	−1.93	−1.93	0.66	−1.93	1.30	1.68	1.07	−1.93
T	0.15	0.66	−1.93	−1.93	1.07	0.66	0.79	0.00	0.00	0.79	−1.93	−1.93	−0.66	−1.93

See Wasserman and Sanderlin, Nature review genetics, 2004 for more details

TF-DNA binding specificity (2)

Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0



Sequence logo

See Wasserman and Sanderlin, Nature review genetics, 2004 for more details

Databases of regulatory sequences

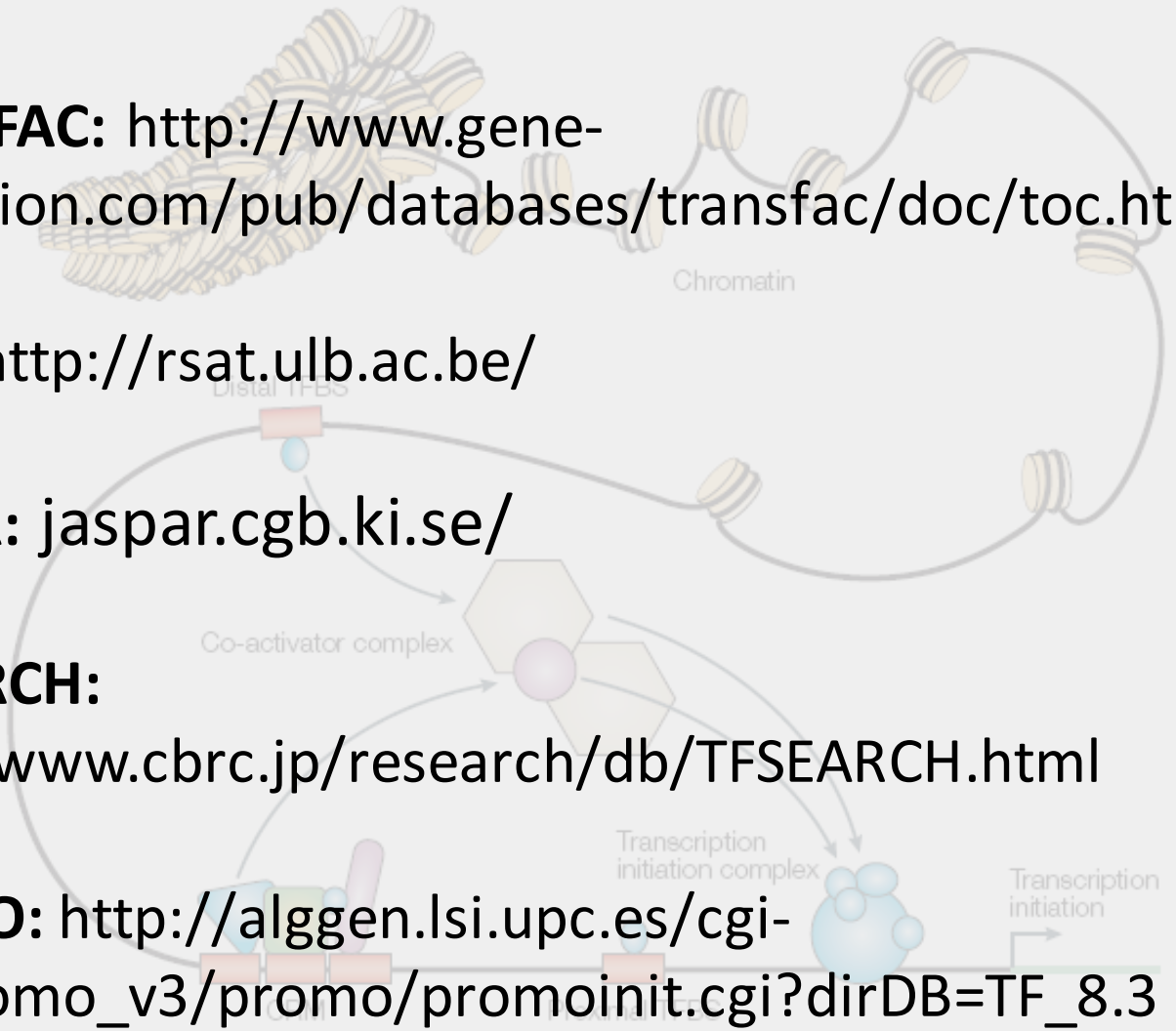
TRANSFAC: http://www.gene-regulation.com/pub/databases/transfac/doc/toc.html

RSAT: http://rsat.ulb.ac.be/

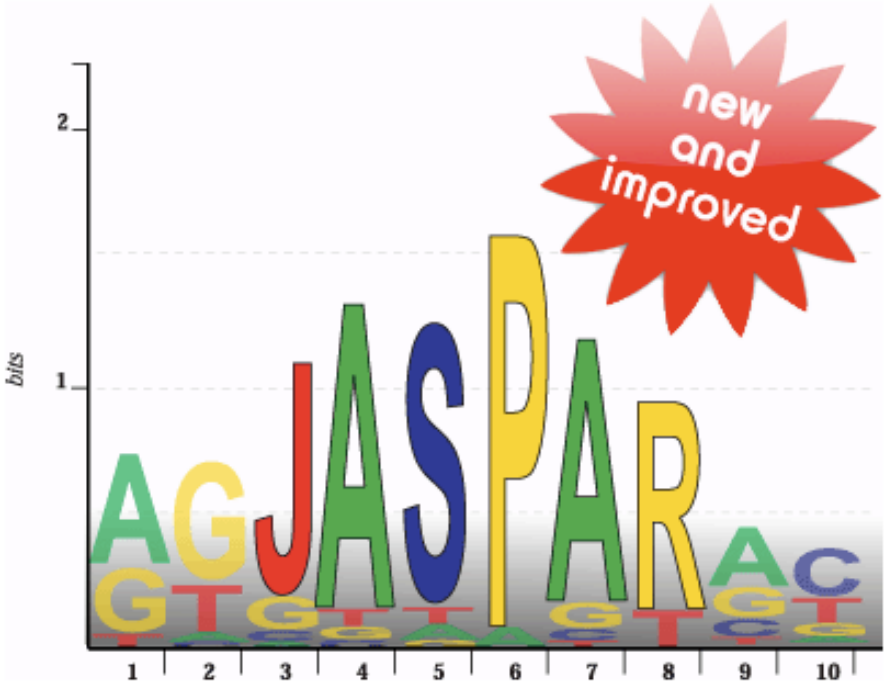
JASPAR: jaspar.cgb.ki.se/

TFSEARCH: http://www.cbrc.jp/research/db/TFSEARCH.html

PROMO: http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3



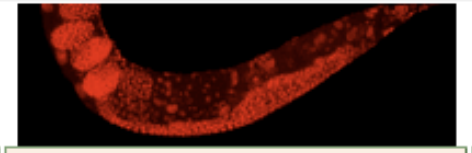
You are using the JASPAR server: jaspar.genereg.net .



The high-quality transcription factor binding profile database

Browse the JASPAR CORE database directly:

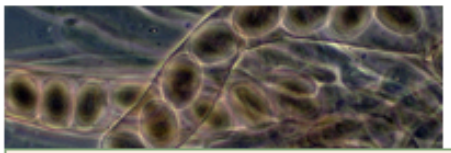
JASPAR CORE Vertebrata



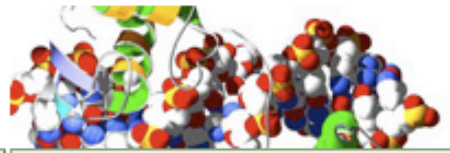
JASPAR CORE Nematoda

JASPAR CORE Insecta

JASPAR CORE Plantae



JASPAR CORE Fungi

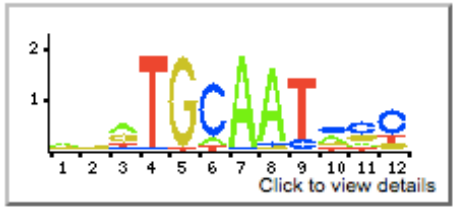
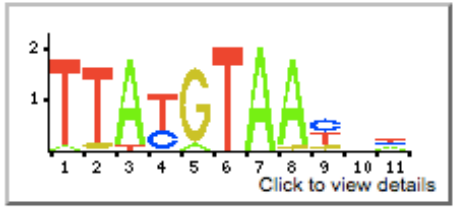


JASPAR CORE by Structural Class

DOCUMENTATION DOWNLOAD CONTACT

Finding potential regulator of gene X

JASPAR matrix models:

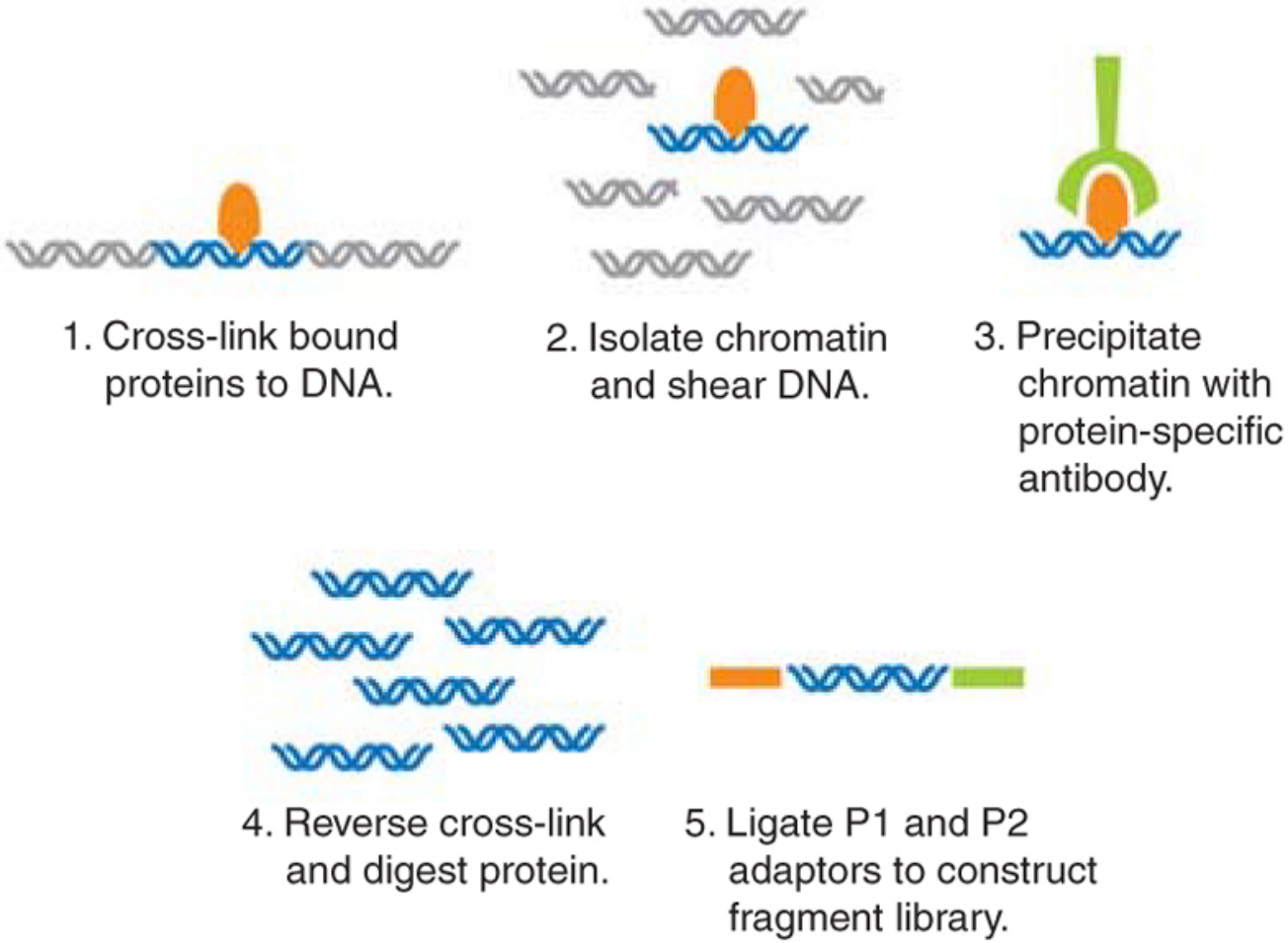
TOGGLE	ID	name	species	class	family	Sequence logo
☑	MA0004.1	Arnt	Mus musculus	Basic helix-loop-helix factors (bHLH)	PAS domain factors	 Click to view details
☑	MA0006.1	Ahr::Arnt	Mus musculus	Basic helix-loop-helix factors (bHLH)::Basic helix-loop-helix factors (bHLH)	PAS domain factors::PAS domain factors	 Click to view details
☑	MA0019.1	Ddit3::Cebpa	Rattus norvegicus	Basic leucine zipper factors (bZIP)::Basic leucine zipper factors (bZIP)	C/EBP-related::C/EBP-related	 Click to view details
☑	MA0025.1	NFIL3	Homo sapiens	Basic leucine zipper factors (bZIP)	C/EBP-related	 Click to view details

Finding potential regulator of gene X

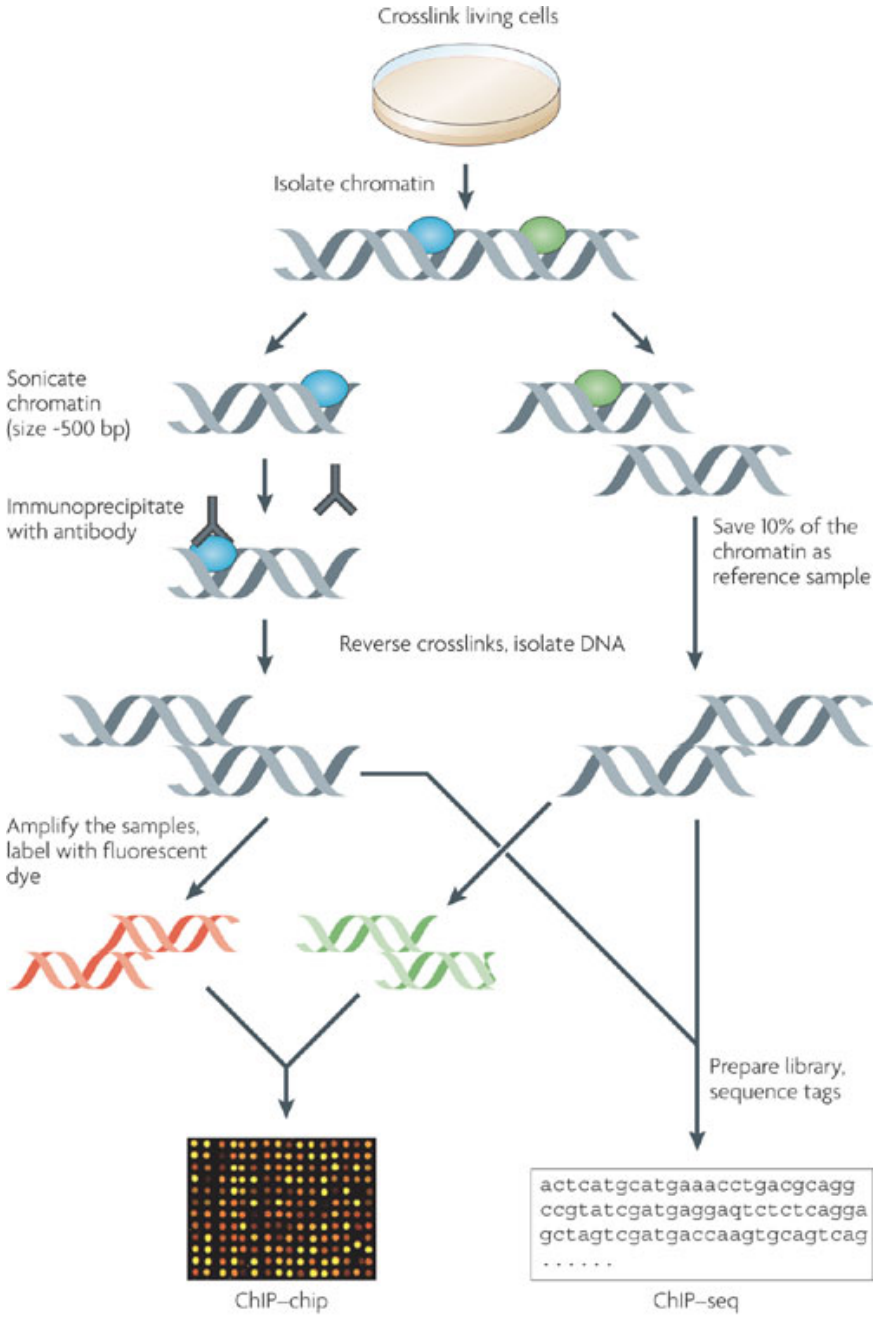
11374 putative sites were predicted with these settings (80%) in sequence named 13

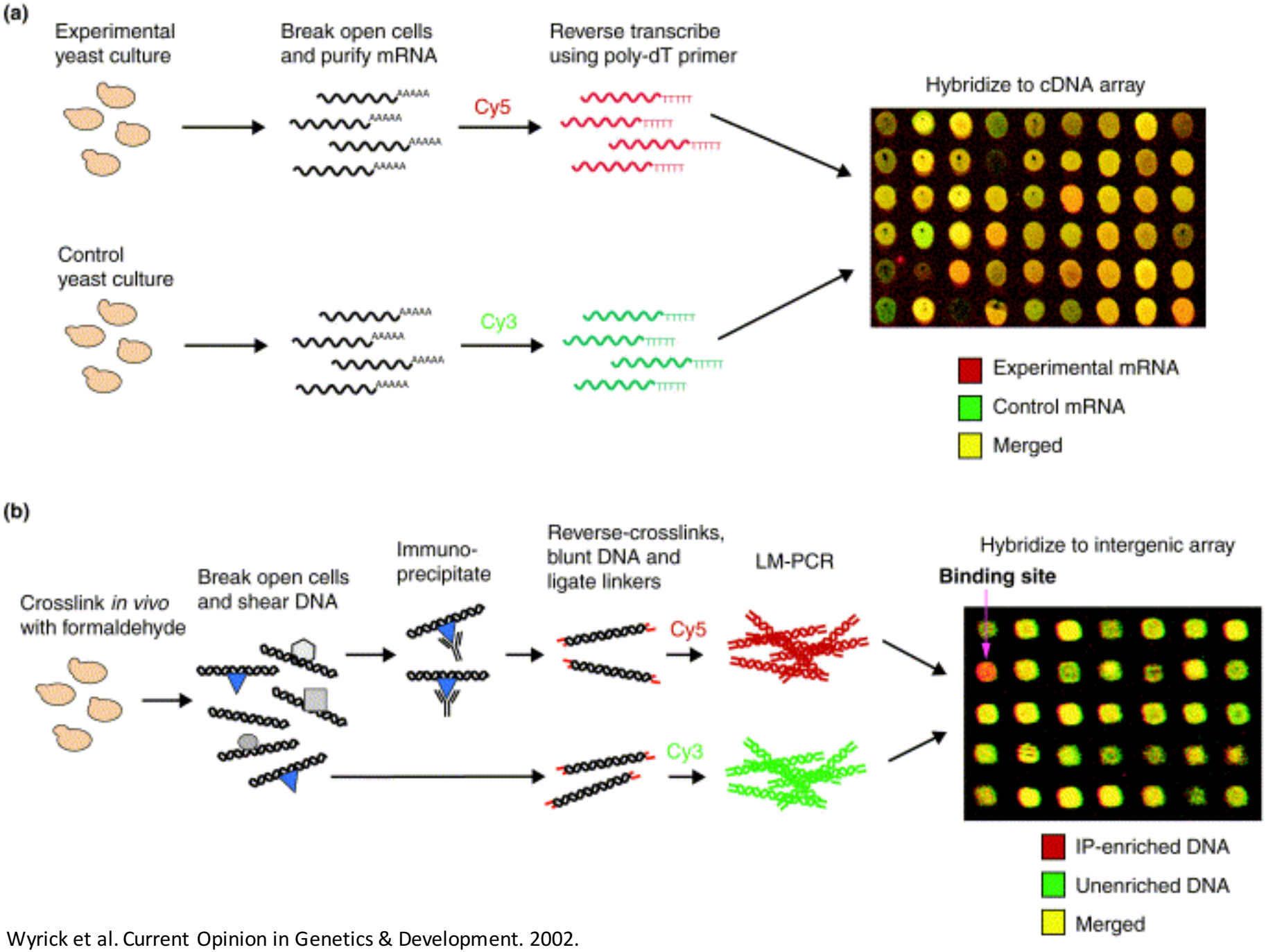
Model ID	Model name	Score	Relative score	Start	End	Strand	predicted site sequence
MA0476.1	FOS	1.335	0.806743455751239	1	11	-1	AATTAGACATC
MA0491.1	JUND	1.005	0.823280060697749	1	11	1	GATGTCTAATT
MA0775.1	MEIS3	6.075	0.868337860898837	1	8	-1	TAGACATC
MA0655.1	JDP2	5.484	0.8197004202266	2	10	1	ATGTCTAAT
MA0498.2	MEIS1	6.198	0.912644042924807	2	8	-1	TAGACAT
MA0899.1	HOXA10	6.405	0.836971511833828	4	14	-1	GGCAATTAGAC
MA0158.1	HOXA5	5.363	0.841814961828179	4	11	1	GTCTAATT
MA0909.1	HOXD13	4.488	0.804219559485583	4	13	-1	GCAATTAGAC
MA0913.1	Hoxd9	5.940	0.836166081301116	4	13	-1	GCAATTAGAC
MA0634.1	ALX3	8.462	0.897808355737197	5	14	-1	GGCAATTAGA
MA0634.1	ALX3	5.348	0.815185380813353	5	14	1	TCTAATTGCC
MA0151.1	Arid3a	5.399	0.826577864606885	5	10	-1	ATTAGA
MA0877.1	Barhl1	9.577	0.971239014736542	5	14	1	TCTAATTGCC
MA0635.1	BARHL2	6.366	0.889002993550392	5	14	1	TCTAATTGCC
MA0878.1	CDX1	5.650	0.84325666837775	5	13	-1	GCAATTAGA
MA0879.1	Dlx1	7.972	0.907268968995571	5	14	-1	GGCAATTAGA
MA0612.1	EMX1	5.726	0.824004796979857	5	14	-1	GGCAATTAGA
MA0612.1	EMX1	9.241	0.922767950639343	5	14	1	TCTAATTGCC

Chromatin immunoprecipitation (ChIP)



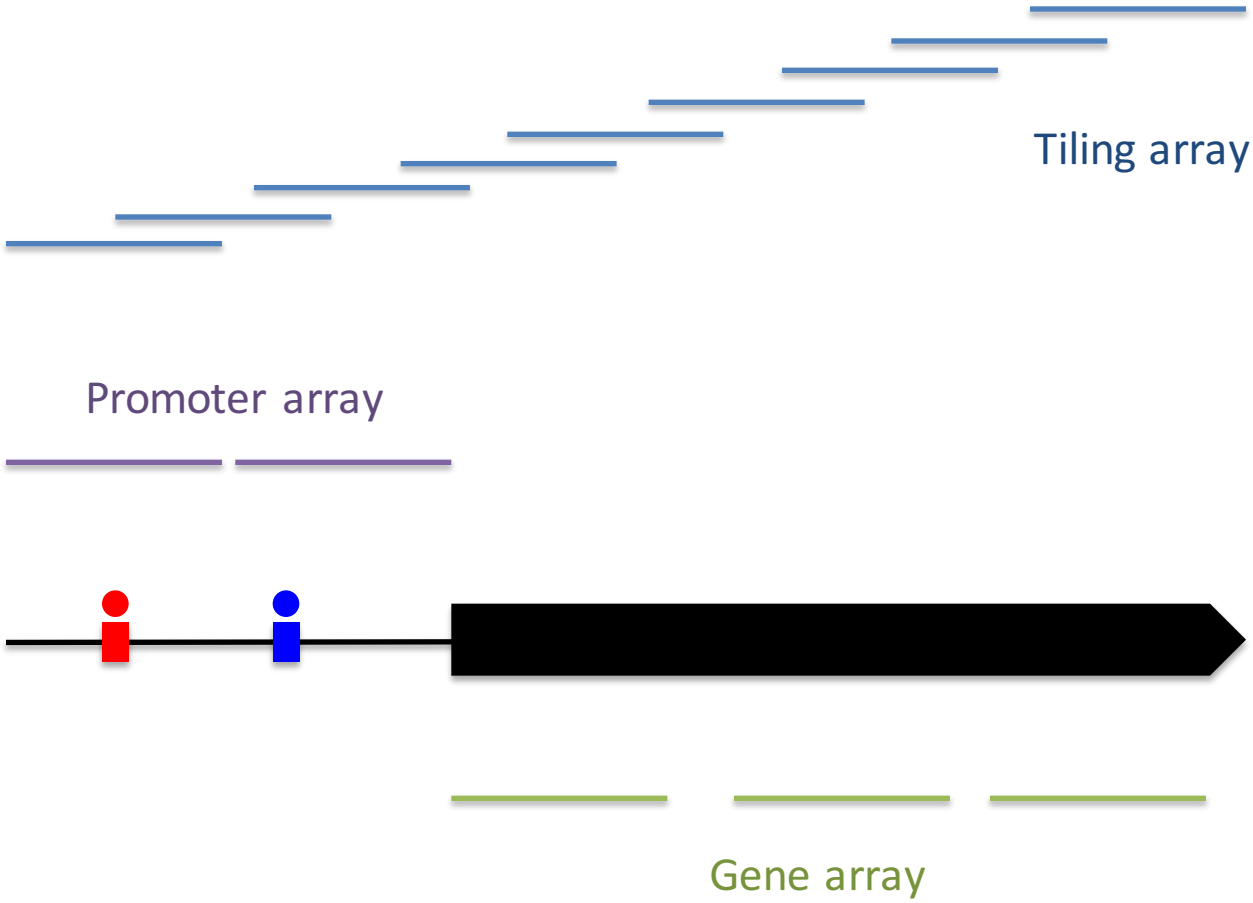
http://www.nature.com/nmeth/journal/v6/n4/fig_tab/nmeth.f.247_F1.html





Wyrick et al. Current Opinion in Genetics & Development. 2002.

Arrays for ChIP-chip or Expression?



Chromatin

Immuno**P**recipitation –

(next-generation)

sequencing

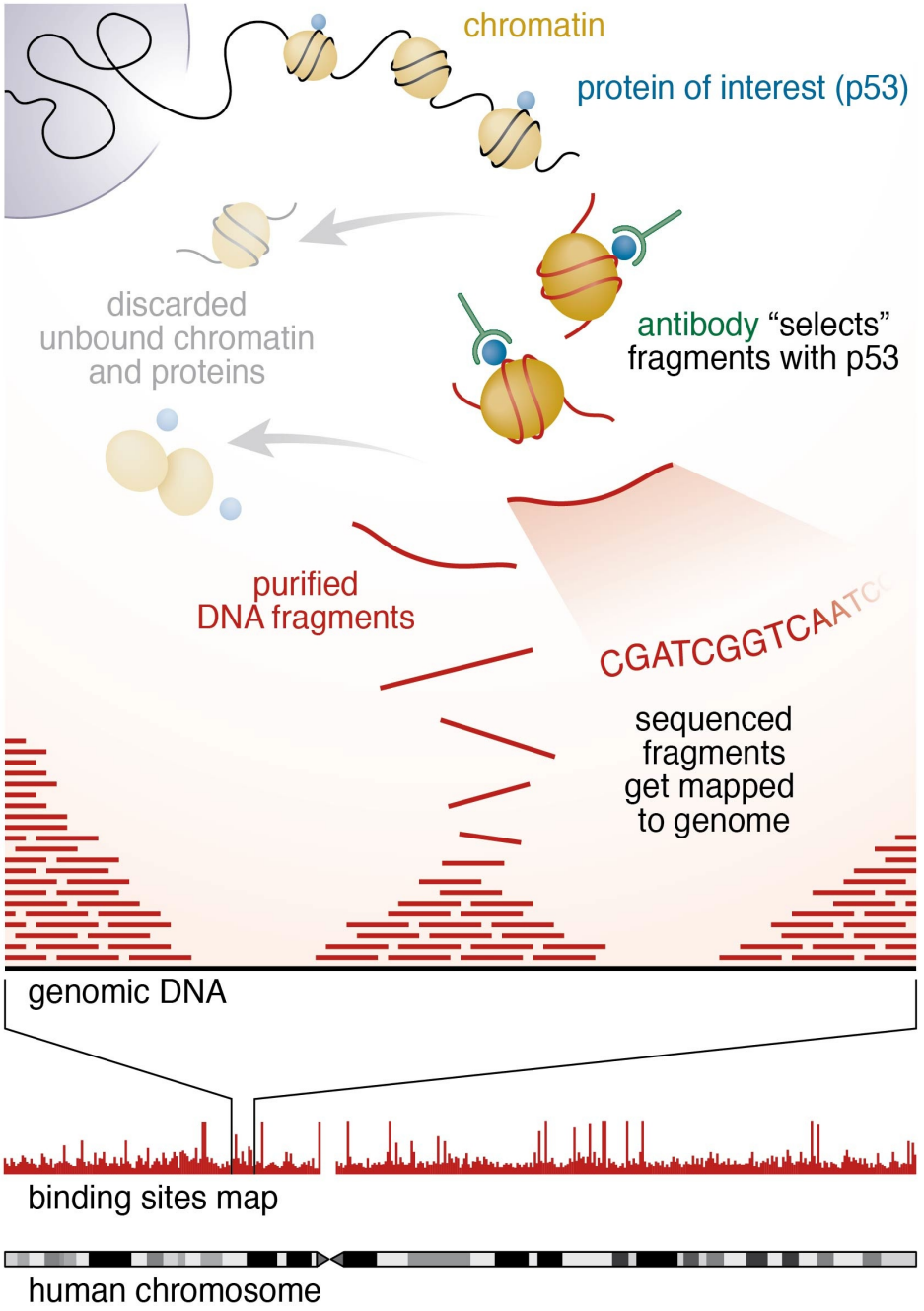


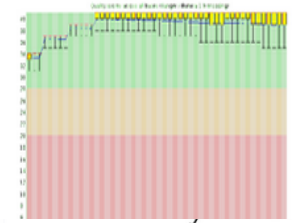
Image from: https://www.bnl.gov/newsroom/news.php?a=111351

ChIP-seq data analysis

read data (FastQ)
rawdata

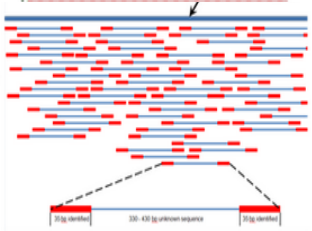
Raw data (sequence with quality scores)

Quality control
fastQC



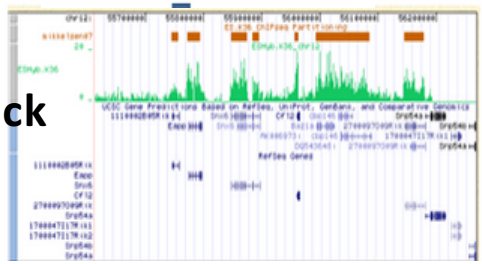
Check quality of reads

Reads mapping
bowtie



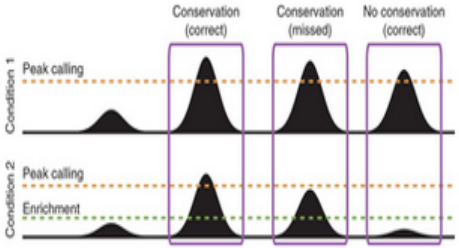
Map and align reads to genome

Build coverage track
perl



Visualize mapped data

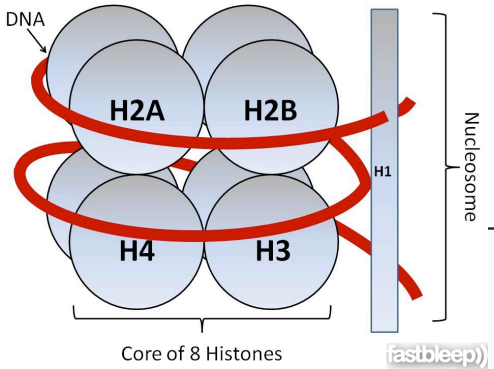
Peak calling
GRange, MACS

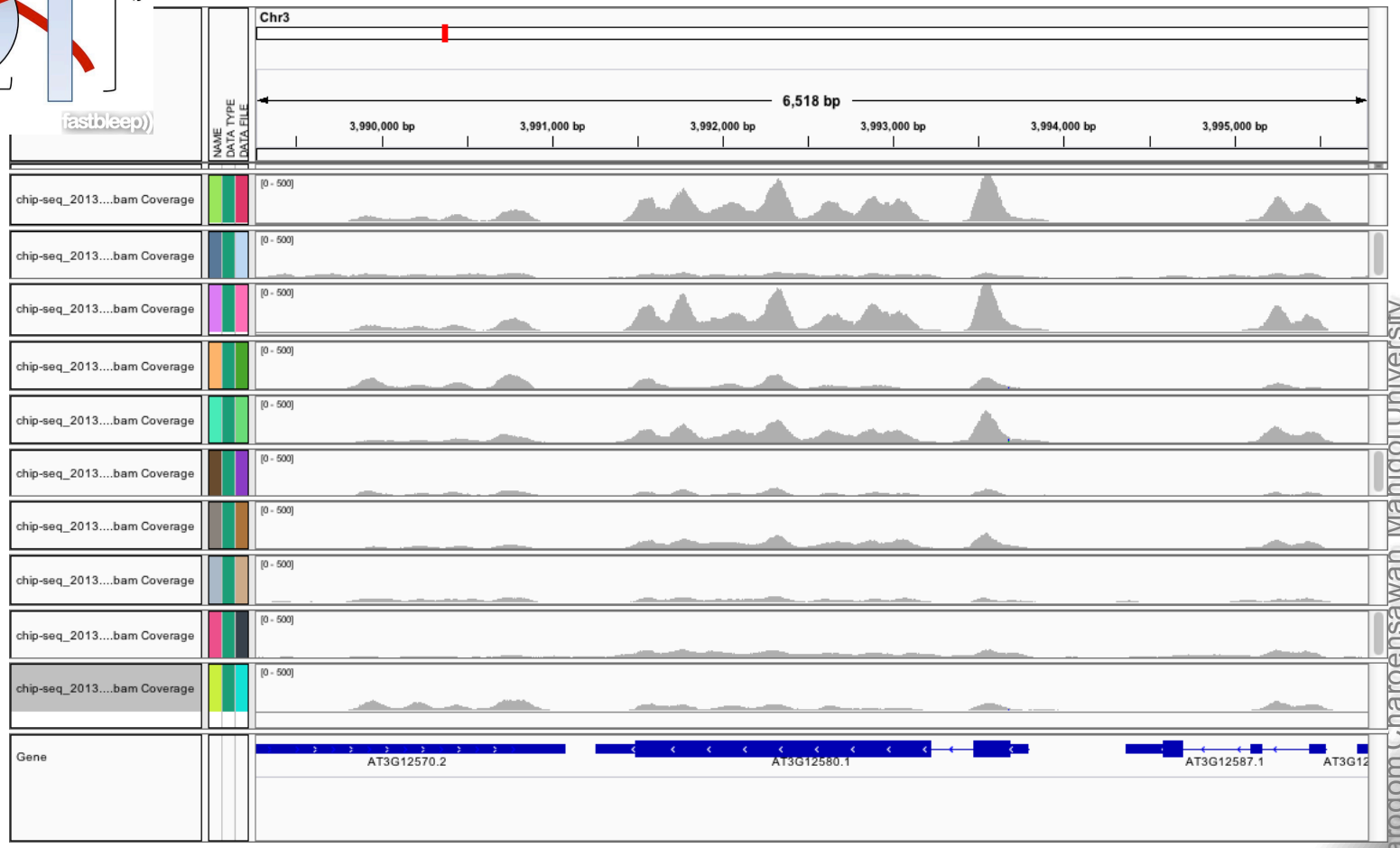


Statistical analysis

Adapted from
http://www4.utsouthwestern.edu/mcdermottlab/NGS/analysis/an
alysis_chipseq.html

Examples of ChIP-data: H2A.Z position at HSP70 promoter





4C hta11::FLAG

4C MNase

12C hta11::FLAG

12C MNase

17C hta11::FLAG

17C MNase

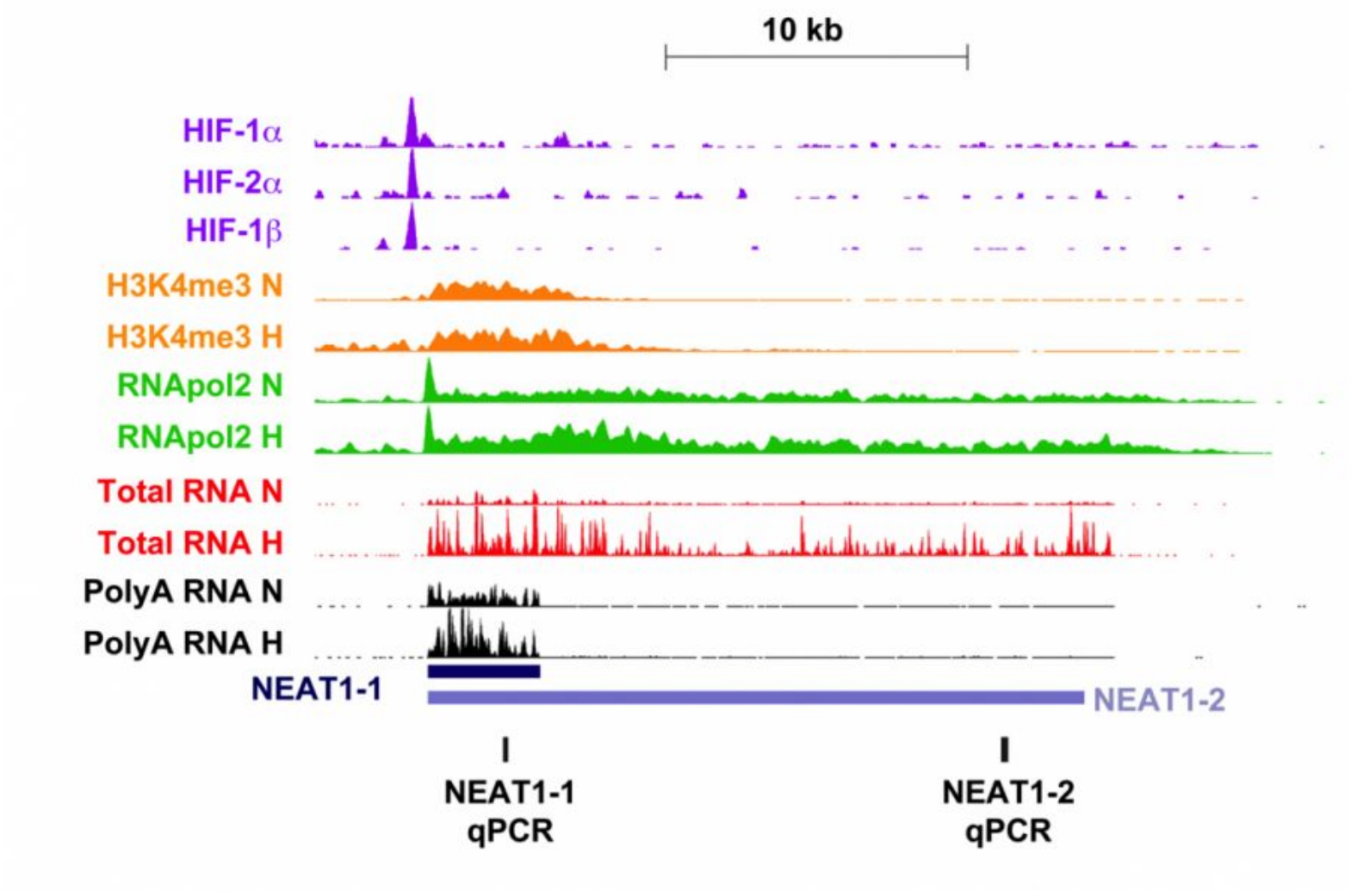
22C hta11::FLAG

22C MNase

27C hta11::FLAG

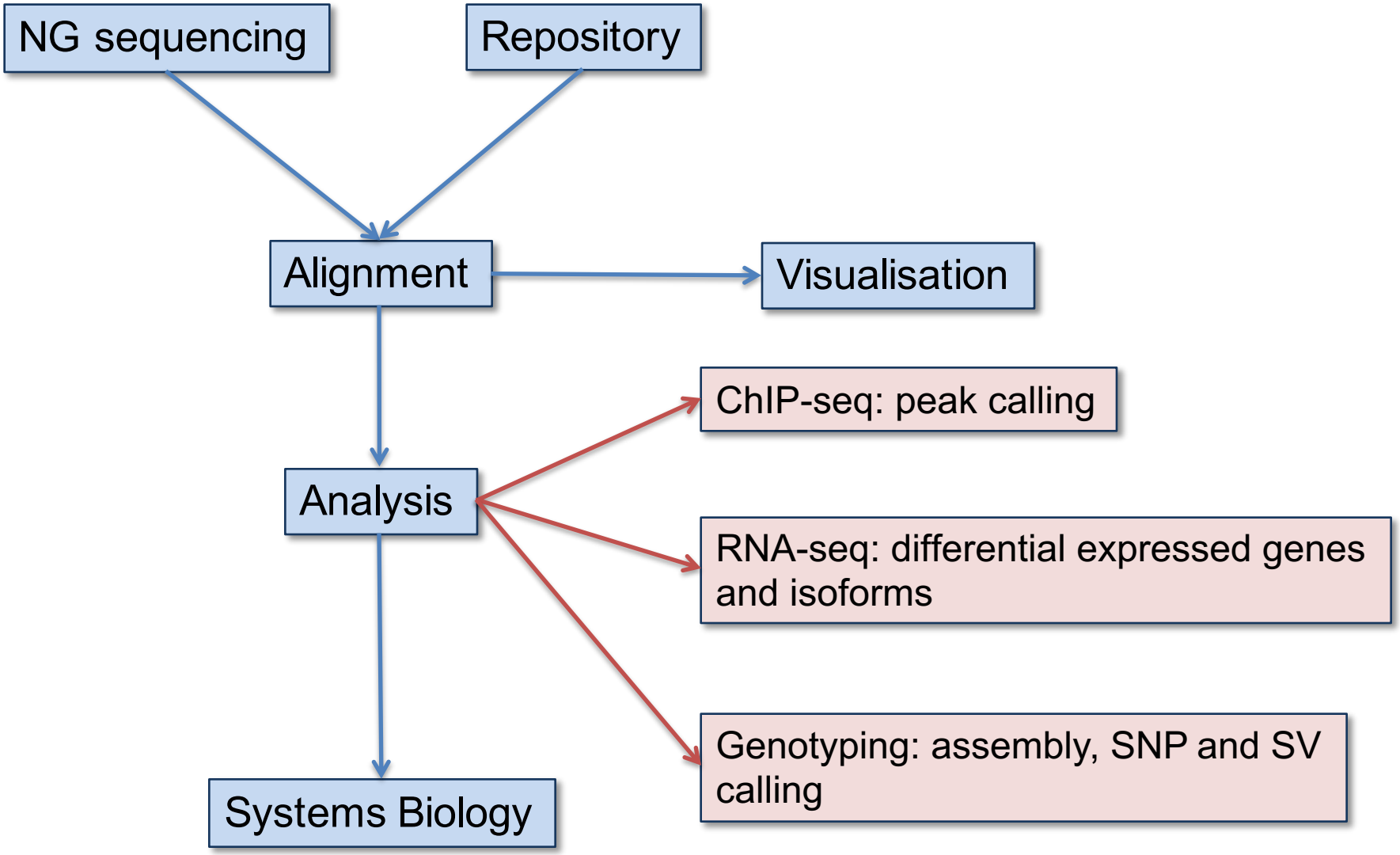
27C MNase

Examples of ChIP-data: transcriptional and epigenetic response to hypoxia in normal and cancer cells

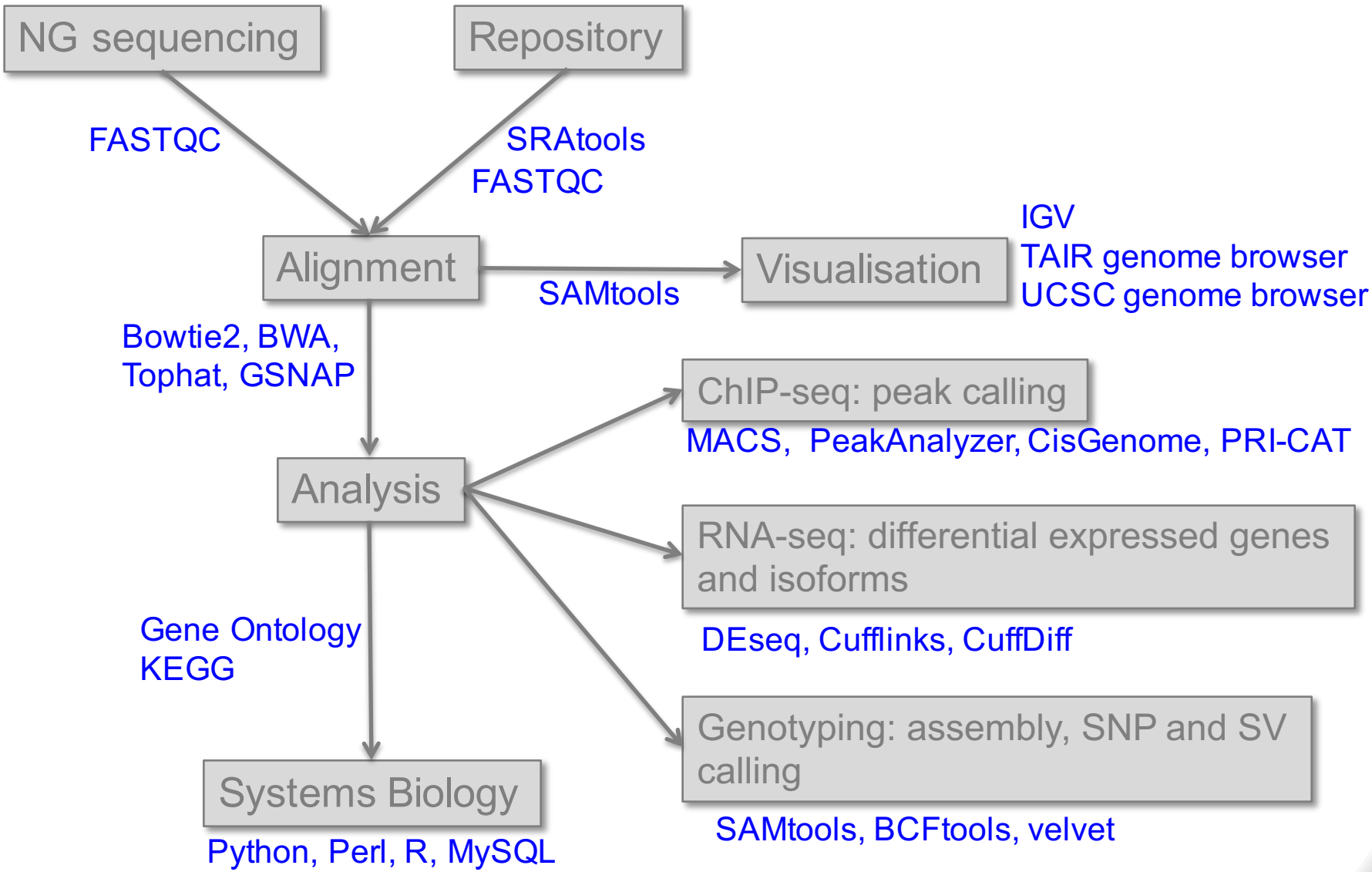


http://www.ndm.ox.ac.uk/principal-investigators/project/modeling-the-transcriptional-and-epigenetic-response-to-hypoxia-in-normal-and-cancer-cells

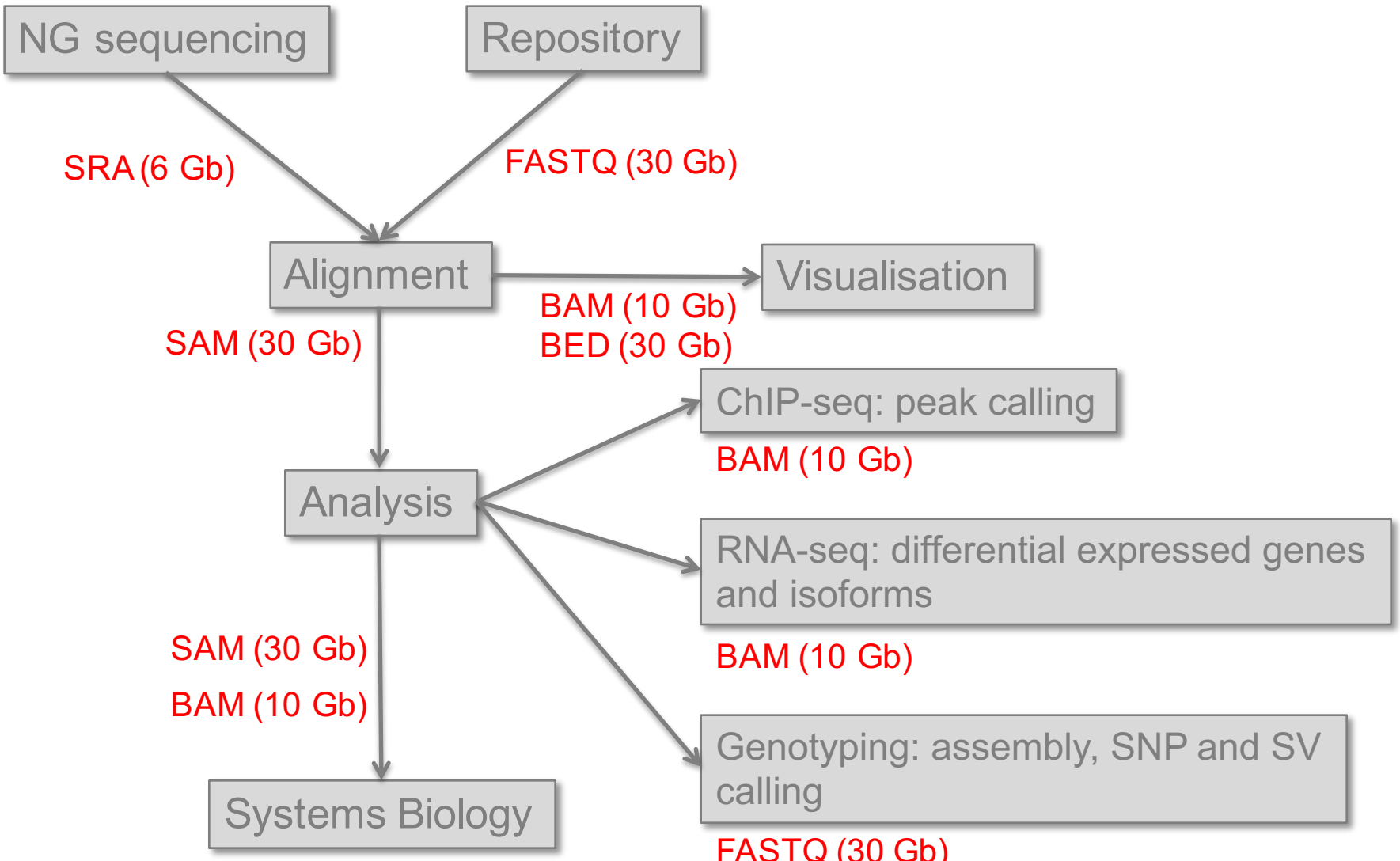
Computational Steps of NGS Data Analyses



Computational Steps of NGS Data Analyses



Computational Steps of NGS Data Analyses



FASTQ and Phred score

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

$$Q = -10 \log_{10} P$$

or

$$P = 10^{\frac{-Q}{10}}$$

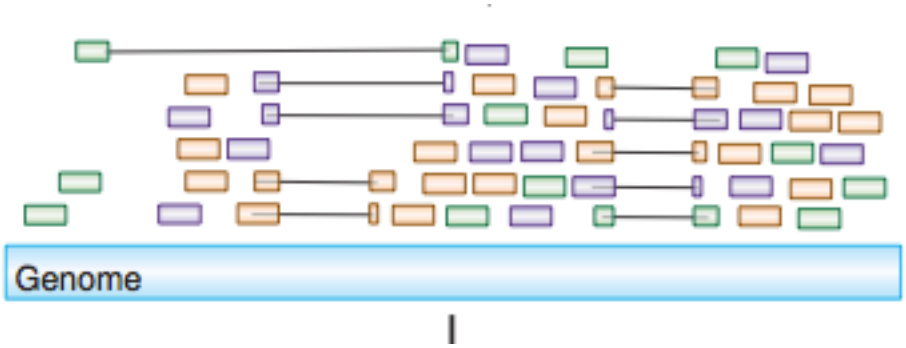
Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Phred score + 33 -> ASCII character

http://en.wikipedia.org/wiki/FASTQ_format

SAM format



mapped short reads

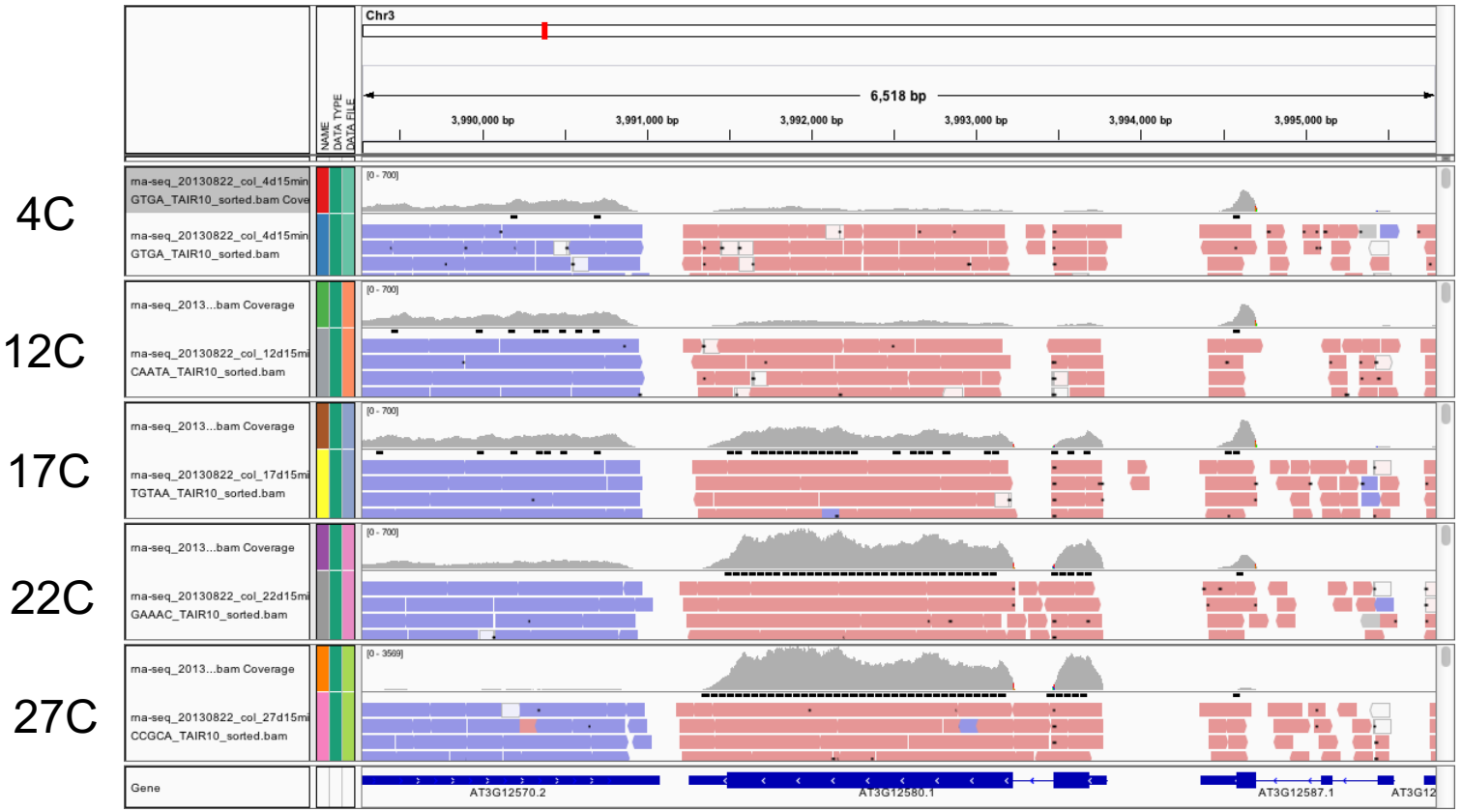
```
1:497:R:-272+13M17D24M 113    1   497    37 37M15
    100338662 0
    CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
    0;==-==9;>>>>>>=>>>>>>>>>>>>=>>>>>>>>>>
    XT:A:U NM:i:0 SM:i:37   AM:i:0 X0:i:1 X1:i:0 XM:i:0
    XO:i:0 XG:i:0 MD:Z:37
```

Hass and Zody 2010; http://genome.sph.umich.edu/wiki/SAM

Outlines

- Transcriptomics: genome-wide transcriptional read-outs (microarray, RNA-seq)
- Regulation of transcription (and transcriptomes)
  - Computational prediction
  - *In vitro*: EMSA, PBM
  - *In vivo:* ChIP-chip, ChIP-seq
- Putting things in perspectives

Examples of RNA-seq data: HSP70 as a temperature sensor

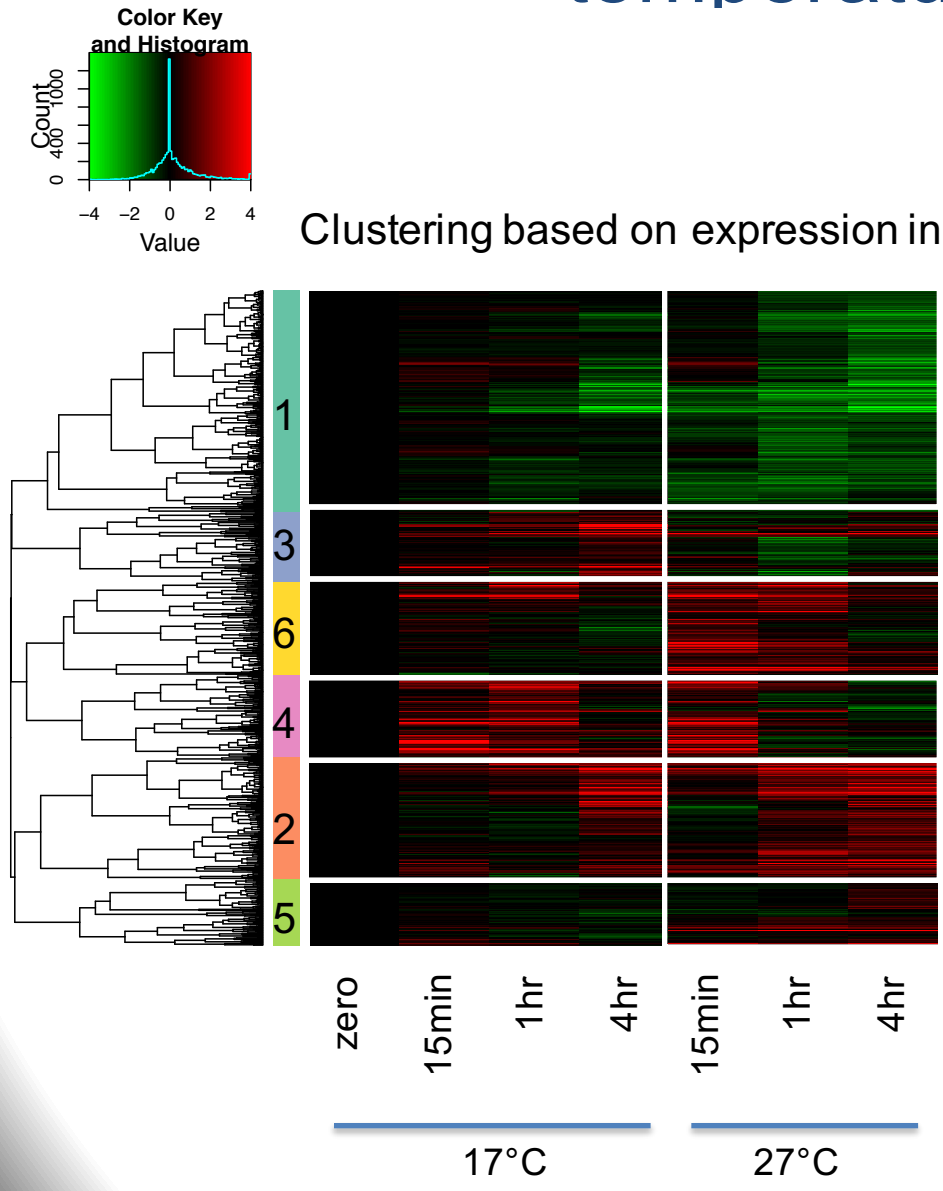


Transcriptional response after 15 mins

(Scale 0-600 for all, except 27C: 0-3500)

With Sandra Cortijo and Phil Wigge

Transcriptome dynamics in response to temperature variation



Cluster 1: 350 genes
GO enrichment in metabolic processes

Cluster 3: 110 genes
No GO enrichment

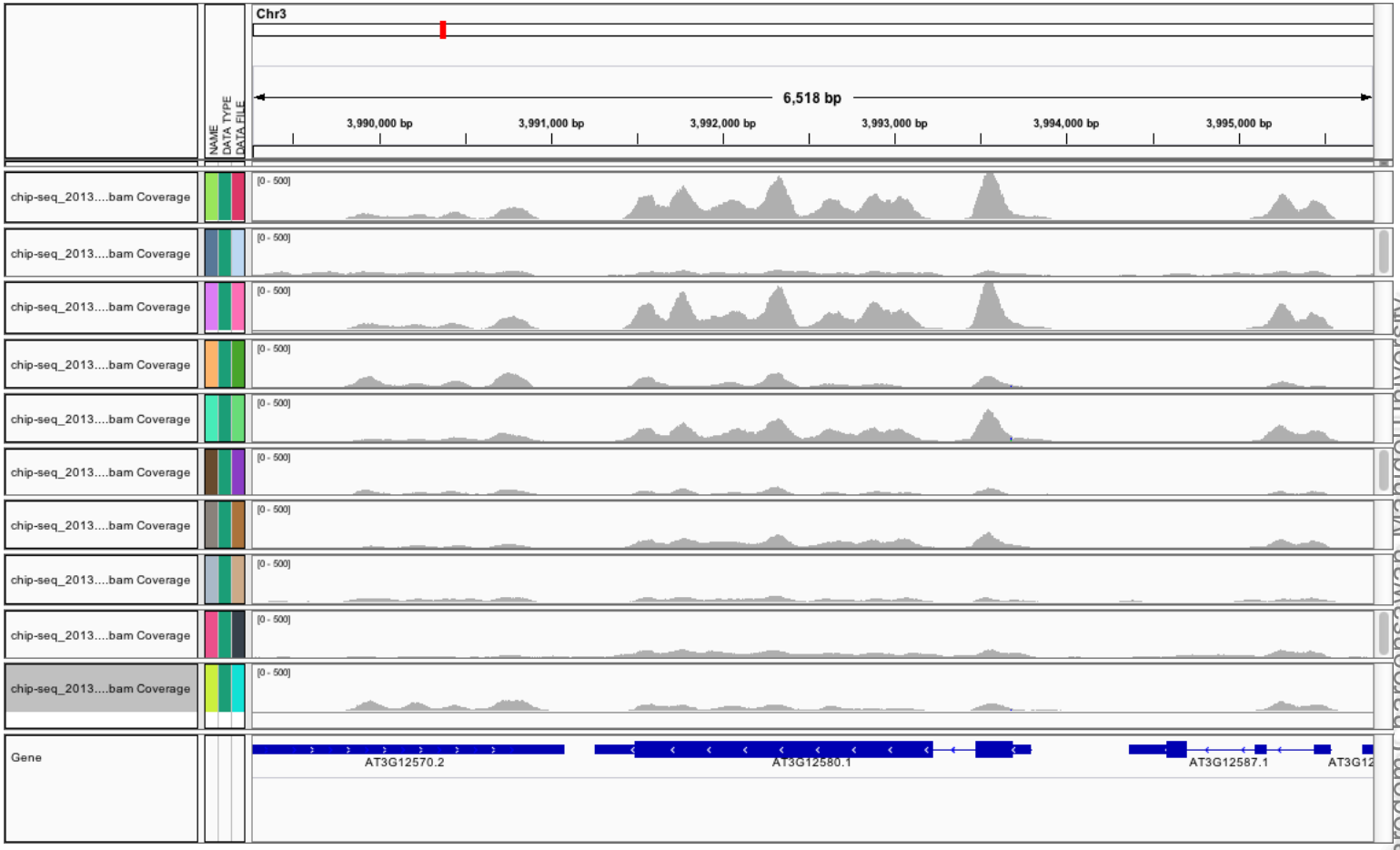
Cluster 6: 147 genes (HSP70) *
GO enrichment in response to heat and protein folding

Cluster4: 130 genes
GO enrichment in regulation of hypersensitive response and immune response

Cluster 2: 193 genes
No GO enrichment

Cluster 5: 105 genes
No GO enrichment

Examples of ChIP-data: H2A.Z position at HSP70 promoter



4C hta11::FLAG

4C MNase

12C hta11::FLAG

12C MNase

17C hta11::FLAG

17C MNase

22C hta11::FLAG

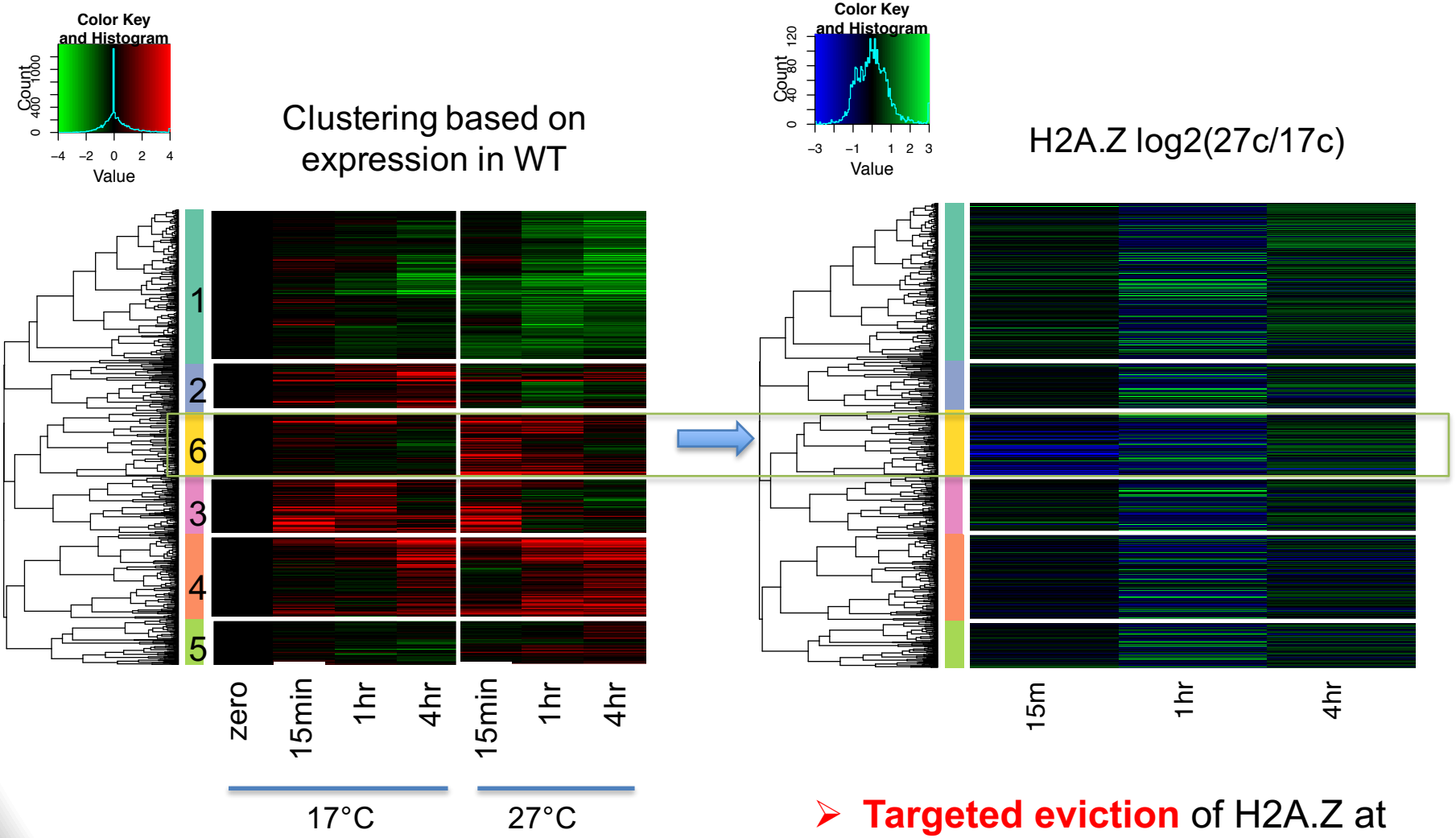
22C MNase

27C hta11::FLAG

27C MNase

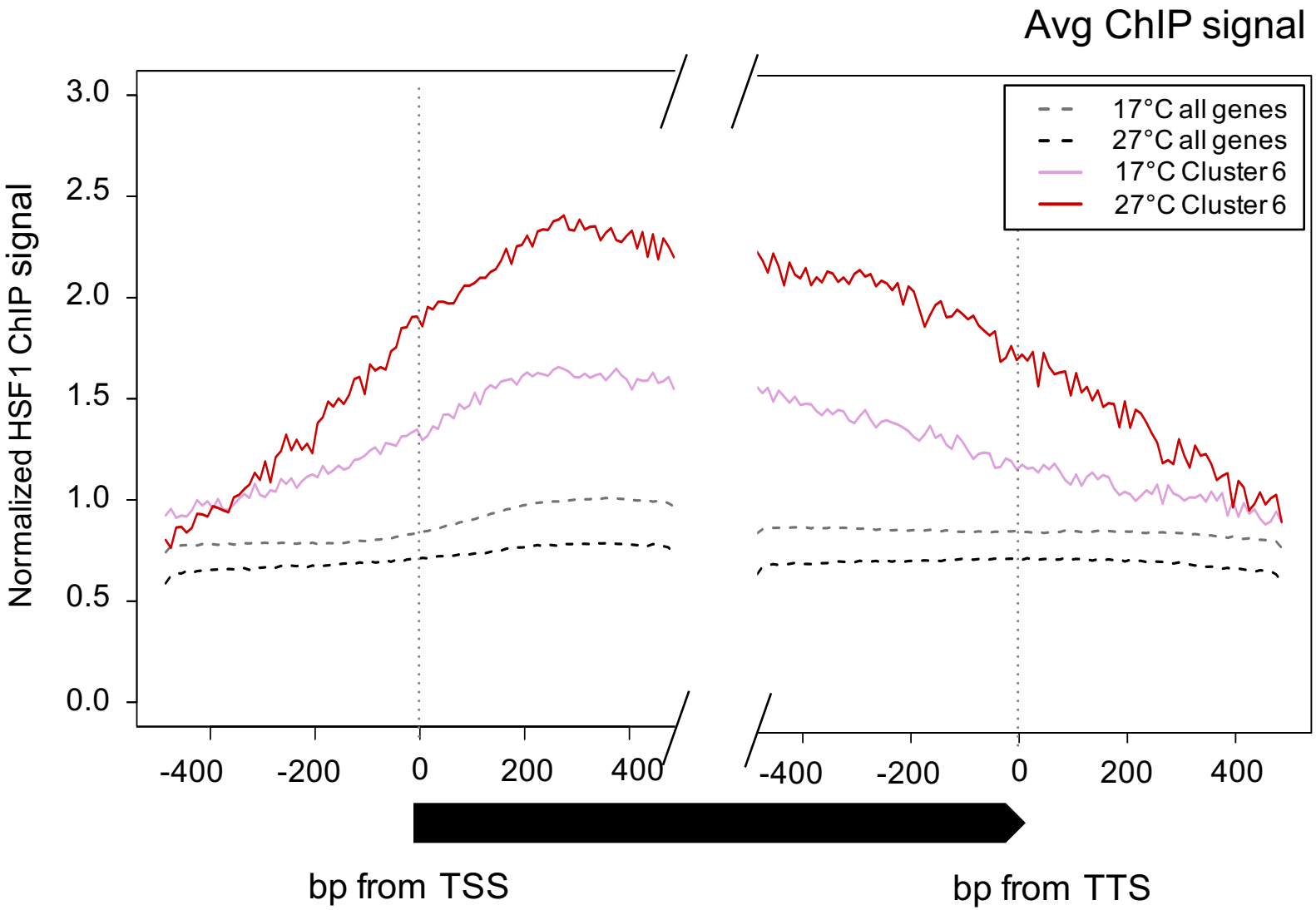
H2AZ occupancy after 1 hr
Scales 0-500 for all

Interplay between H2A.Z and transcriptional changes in response to temperature

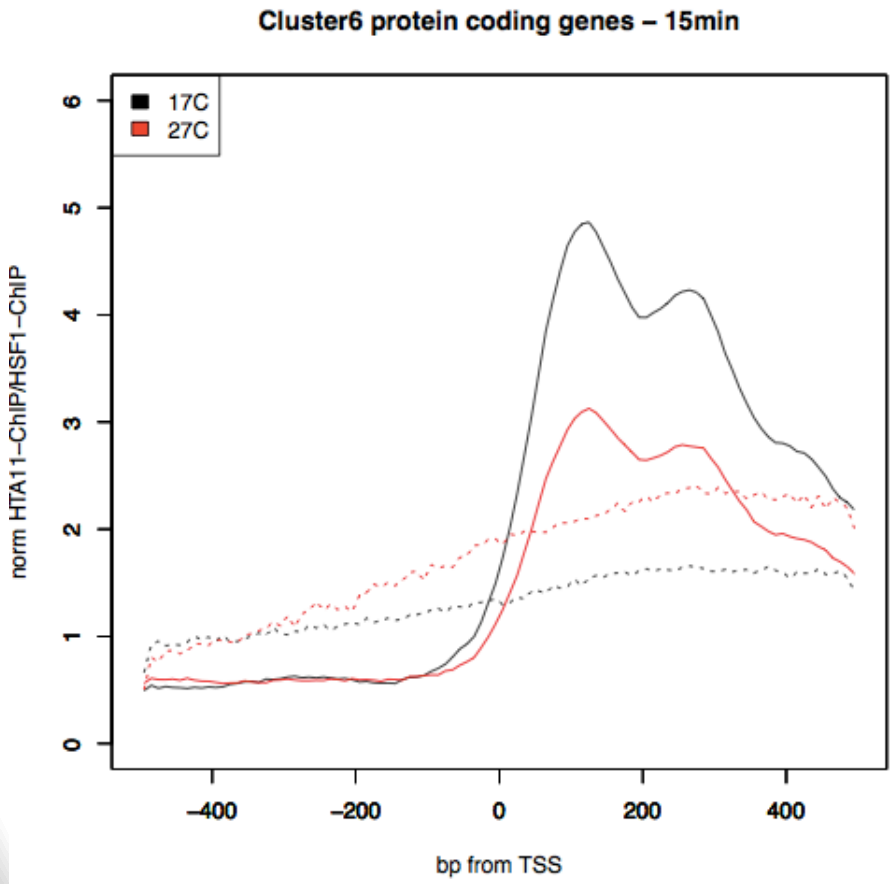


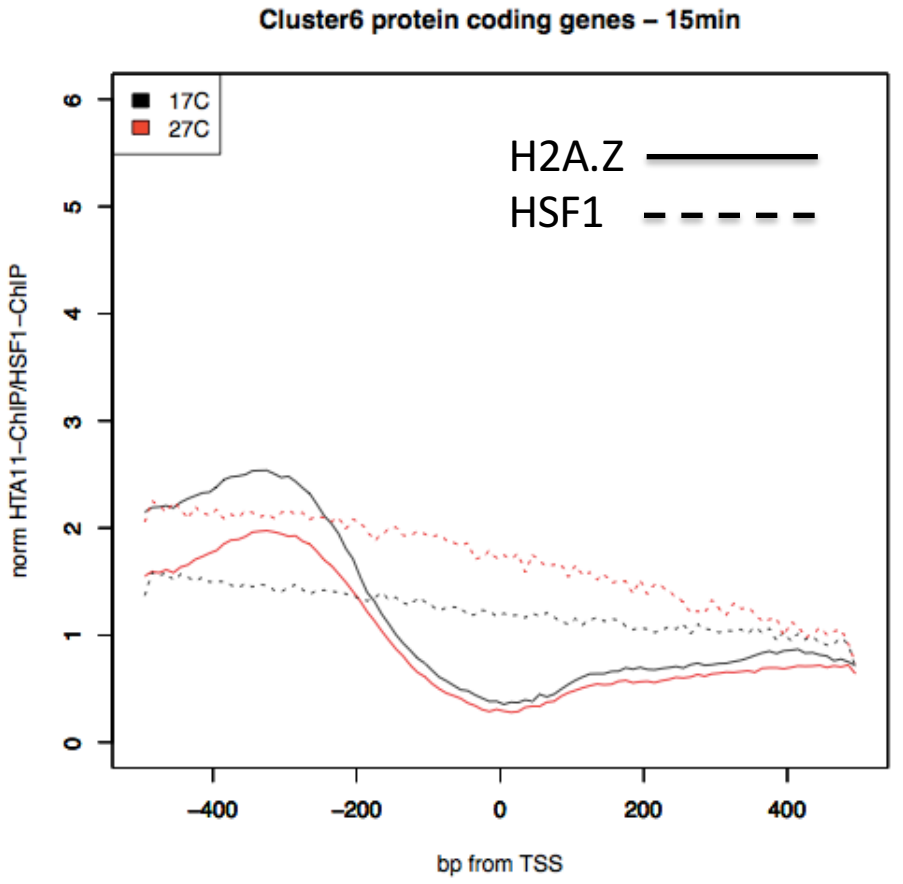
- **Targeted eviction** of H2A.Z at 15min at genes activated by the temperature shift

Increase of **HSF1** binding at highly expressed genes in **Cluster 6**

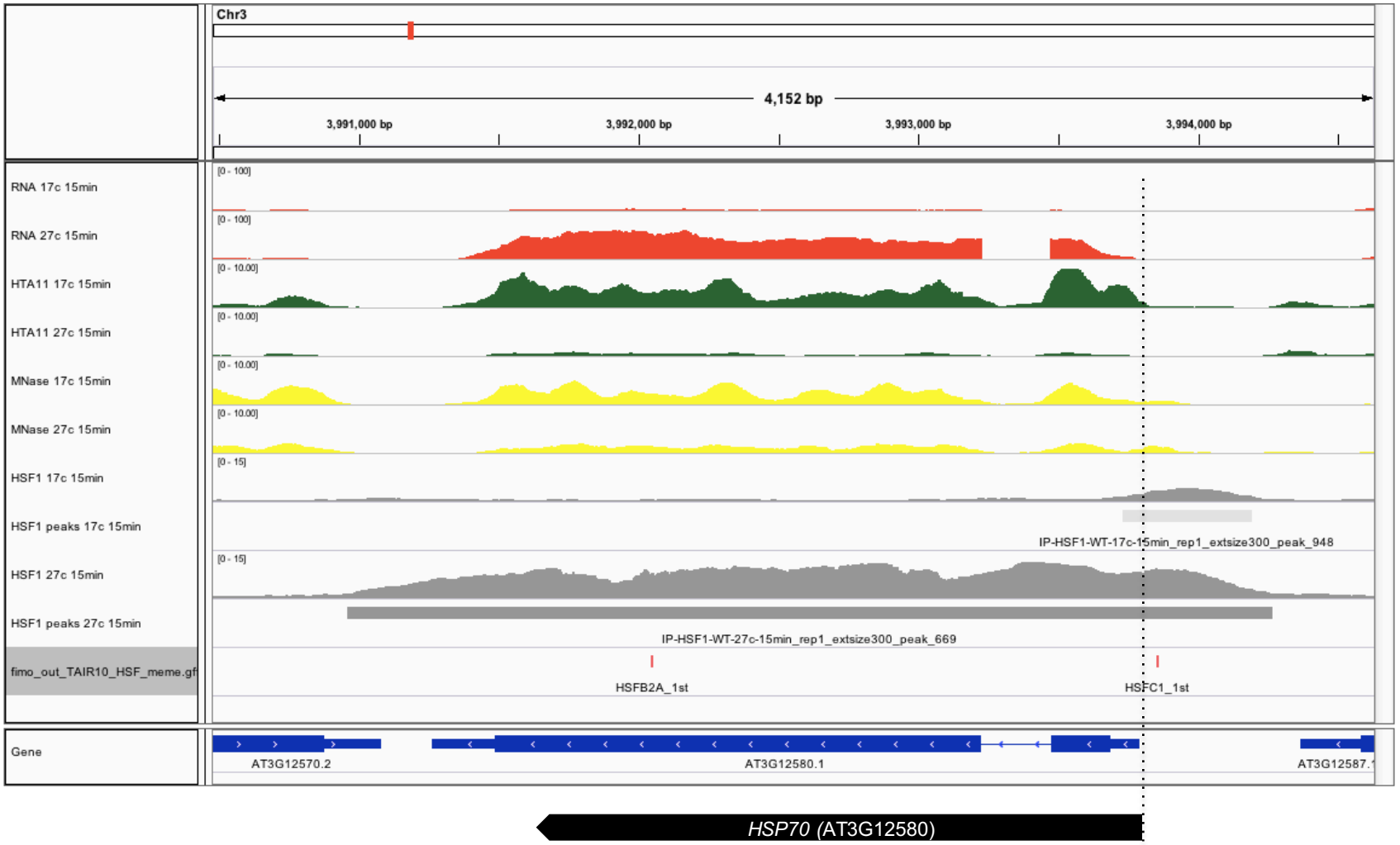


Reduction of nucleosome stability vs' increase of HSF1 binding at highly expressed genes in **Cluster 6**



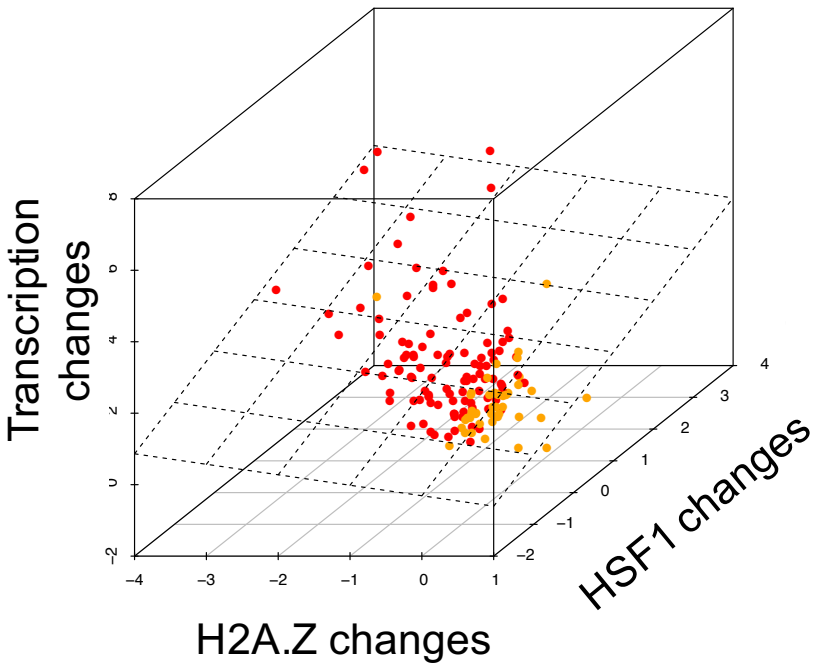


Antagonistic role of HSF1 and H2A.Z



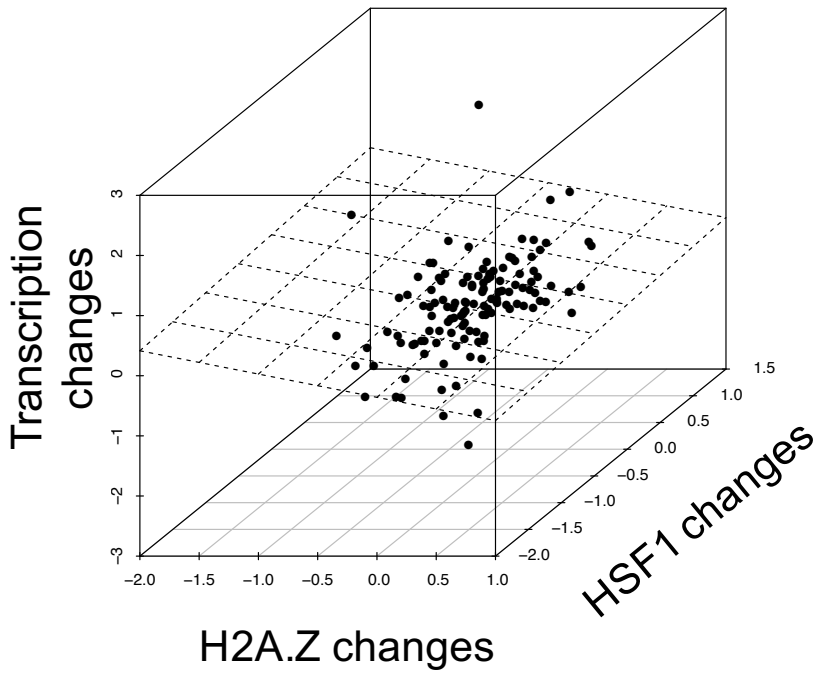
HSF1 binding and H2A.Z eviction are good predictors for transcription of temperature-responsive genes

Cluster 6



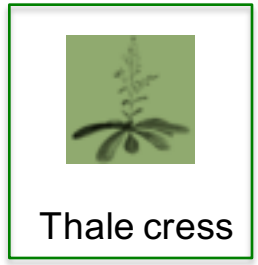
Multiple R-squared: 0.33
P-value: 3.678e-13

Random genes

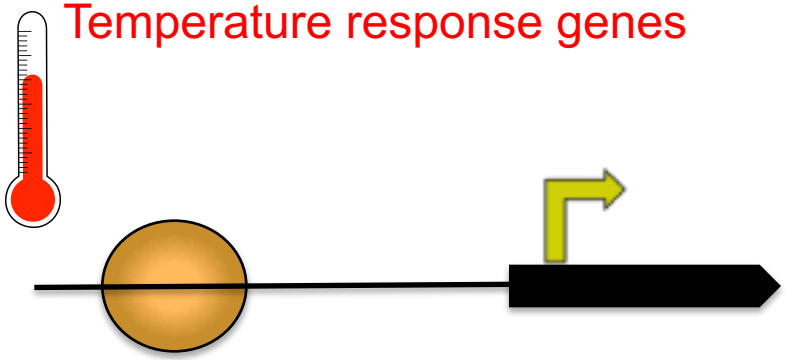


Multiple R-squared: 0.06
P-value: 0.0221

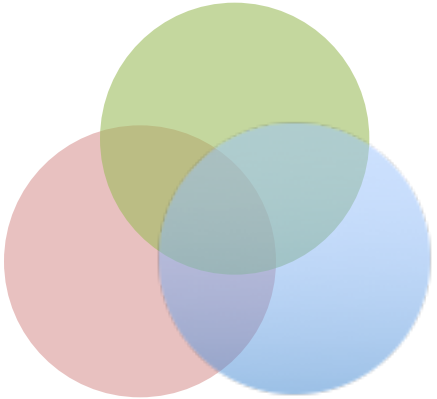
Characterising Transcriptional Architectures of Temperature Response



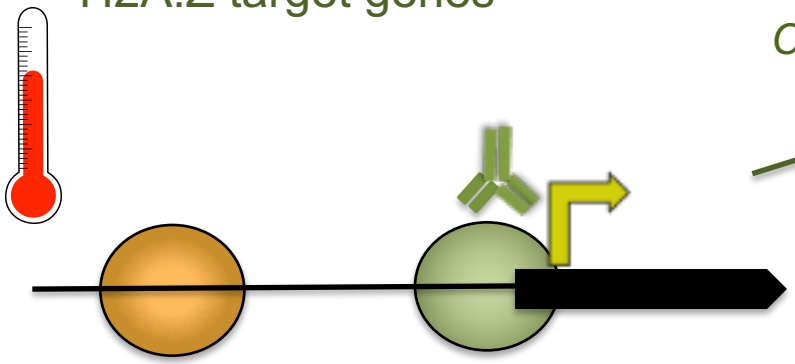
Temperature response genes



RNA-seq



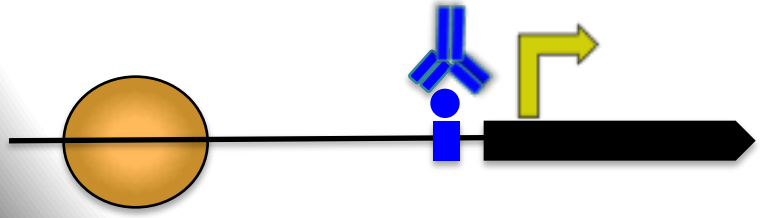
H2A.Z target genes

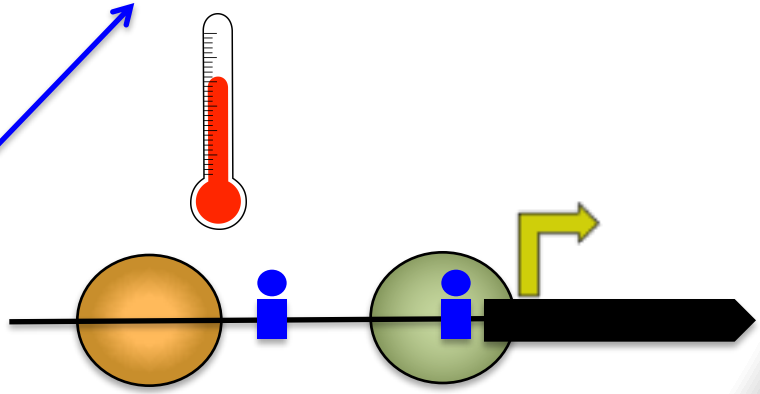


ChIP-seq

ChIP-seq

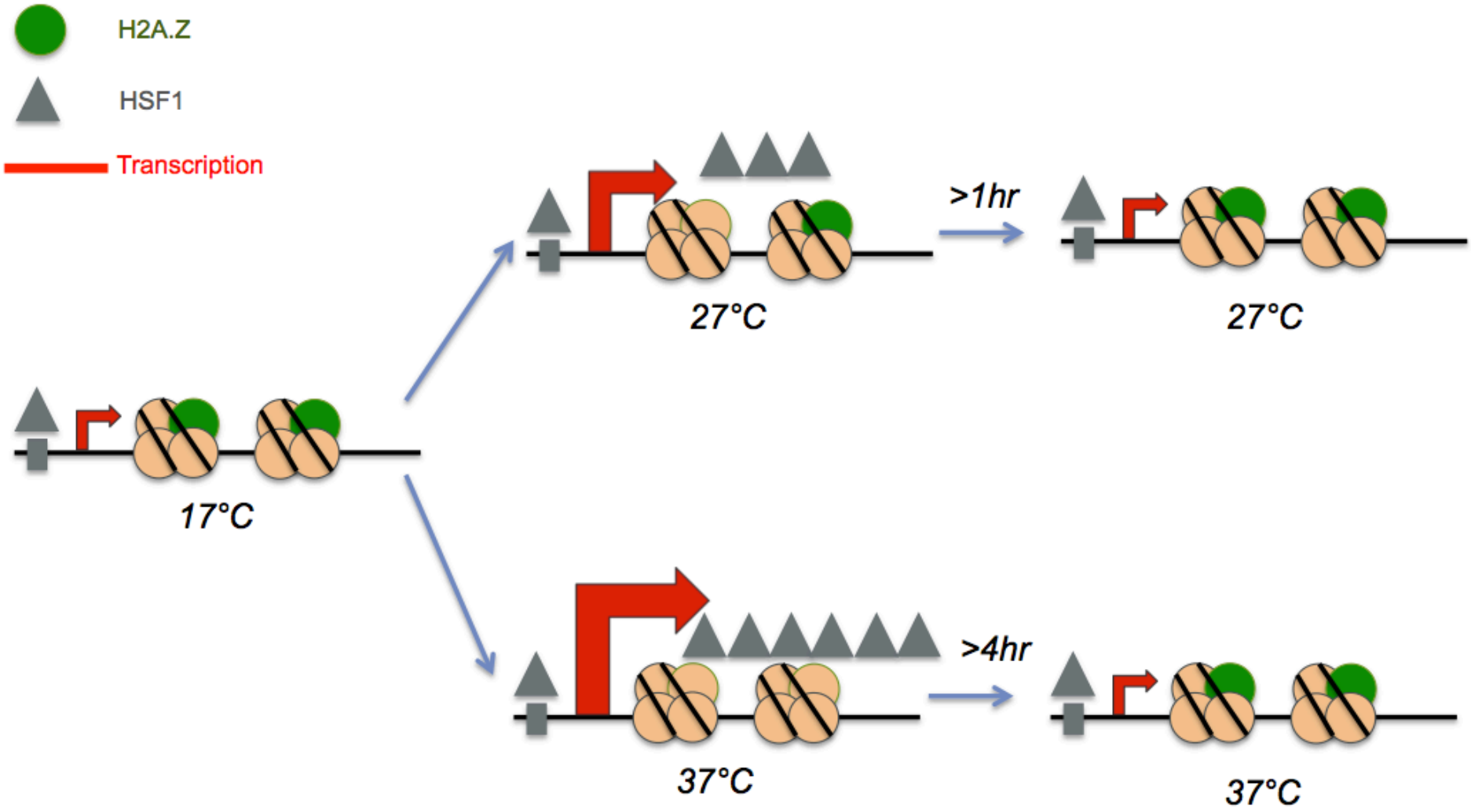
Trans-acting factors' (TFs) target genes





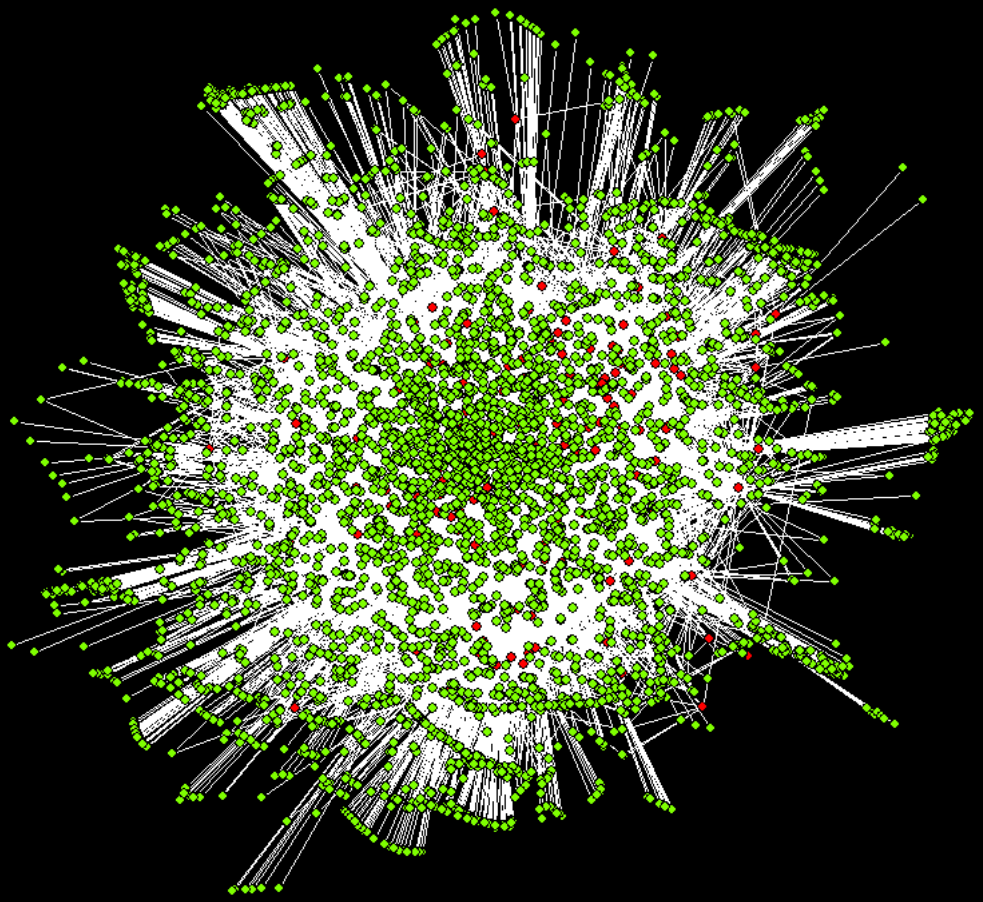
How many? What genes?

How temperature affects gene expression and plant development?



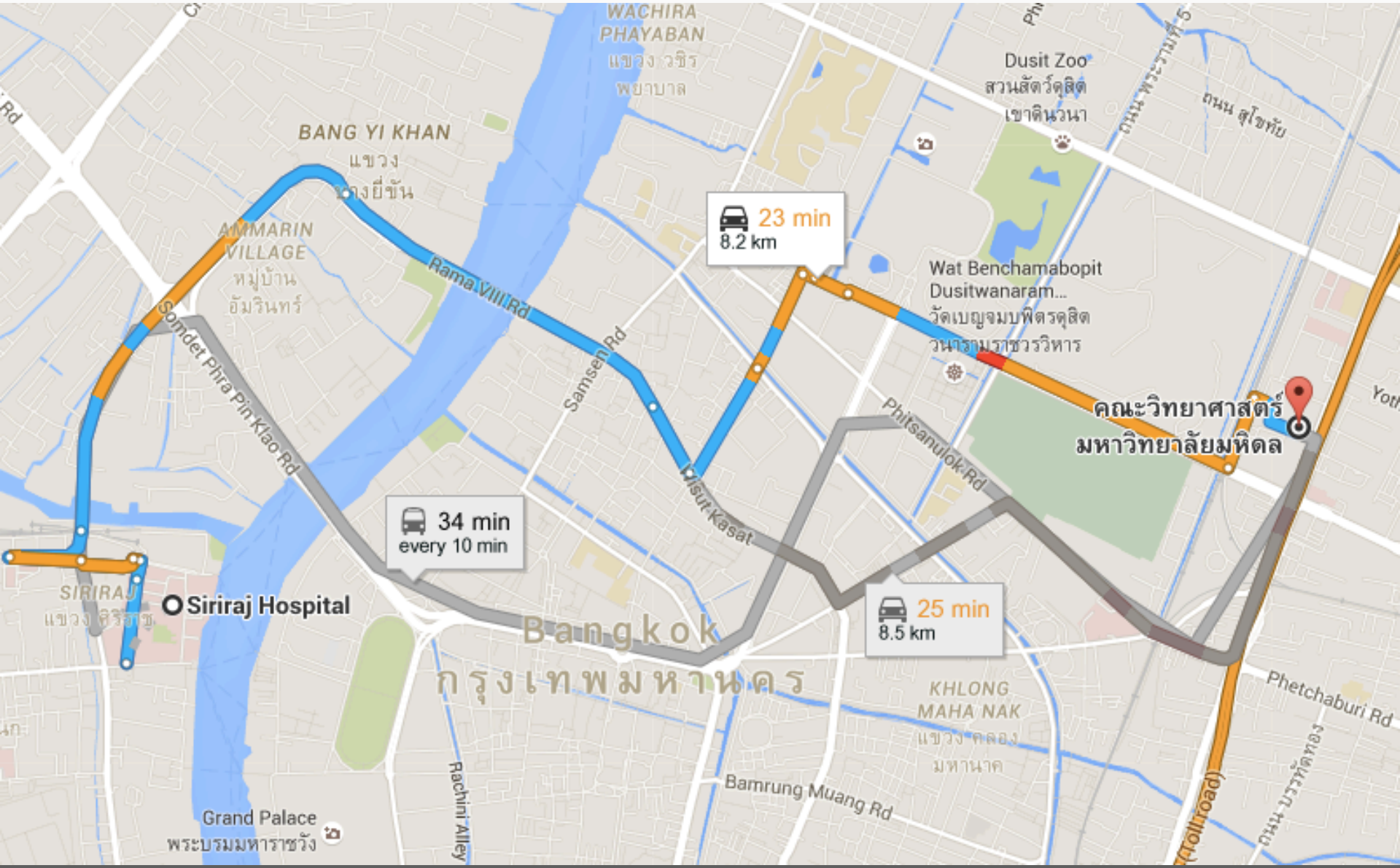
Cortijo* and Charoensawan* *et al. Molecular Plant. 2017*

Network Biology



Slides courtesy of Dr MM Babu, MRC LMB

How to from A to B?



Further readings

- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009 Jan;10(1):57-63

- Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet*. 2012 Dec;13(12):840-52.

- Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods*. 2009 Nov;6(11 Suppl):S22-32. doi: 10.1038/nmeth.1371.

- Hrdlickova R et al. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2017 Jan;8(1). doi: 10.1002/wrna.1364

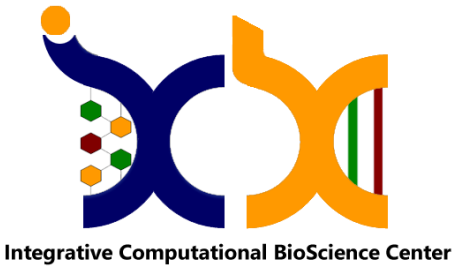
Question? Comment?

Varodom Charoensawan

Department of Biochemistry

Faculty of Science, Mahidol University

varodomc@gmail.com, varodom.cha@mahidol.ac.th



Mahidol University

Wisdom of the Land